Proceedings of the workshop EXTENDING, MAPPING AND FOCUSING THE CRM

ARIADNE

SEVENTH FRAMEWORK PROGRAMME

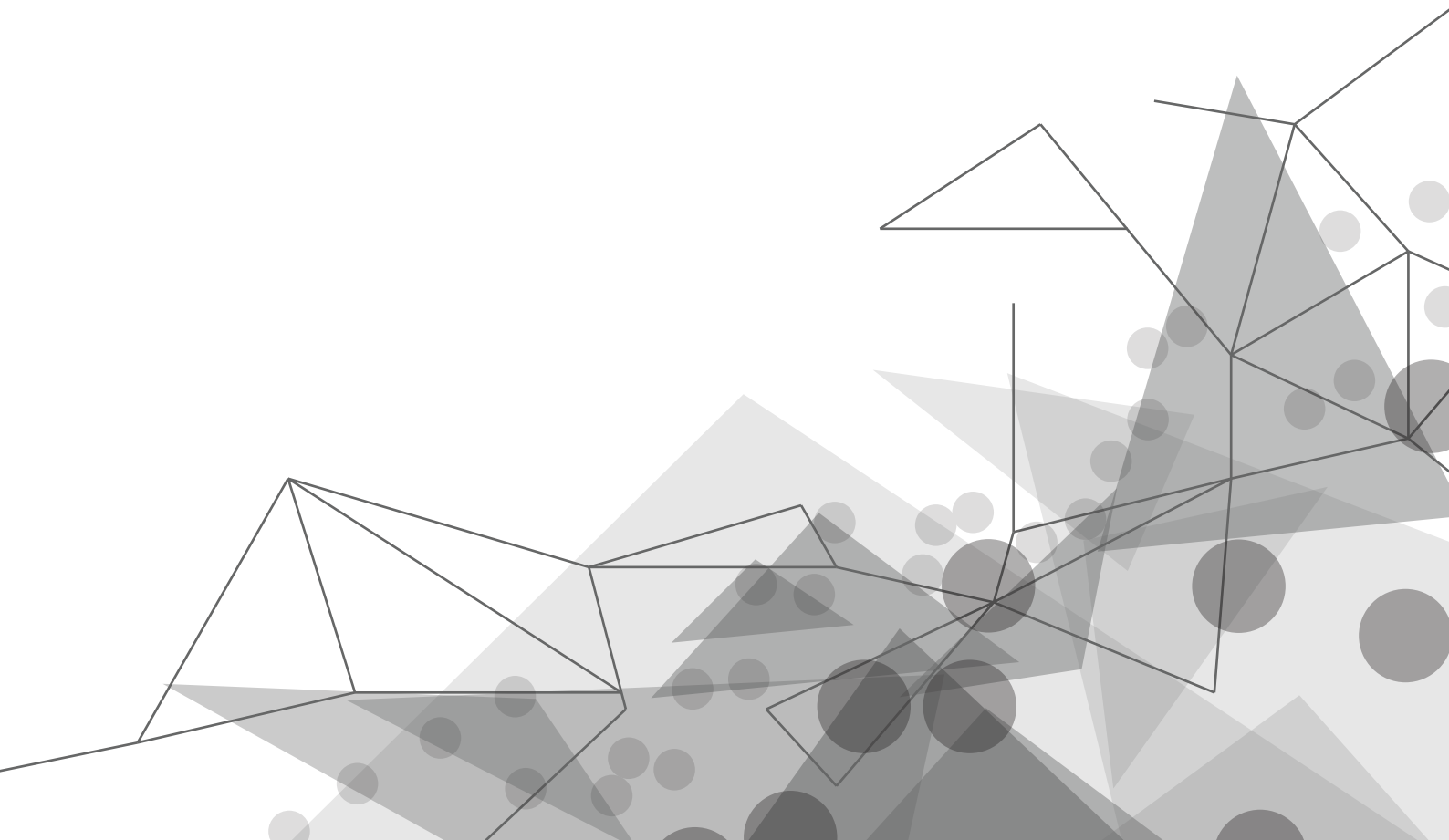
TPDL 2015

14-18/09/2015

Poznań, Poland

Edited by
Paola Ronzino

19th International Conference on Theory and Practice of Digital Libraries



Cite the volume as follows:

Paola Ronzino (ed.): Extending, Mapping and Focusing the CRM 2015. Proceedings Workshop EMF-CRM2015, Poznań, Poland, September 17, 2015, CEUR-WS.org, online CEUR-WS.org/Vol-1656

Vol-1656 urn:nbn:de:0074-1656-8

Cover: Nicola Amico (PIN, Prato)

Preface

The Workshop "Extending, Mapping and Focusing the CRM" was organized in the framework of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL), on the 17th September 2015, in Poznan, Poland. The goal of the workshop was to present, discuss and assess the developments of the CIDOC CRM in the Cultural Heritage domain. The importance of CRM in heritage-related digital libraries and dataset infrastructures is confirmed by the increasing number of cultural heritage institutions and research projects adopting CRM and its extensions to foster datasets interoperability. The CfP was open to scholars contributing to the development of CRM extensions and applications as well as to researchers who are currently using and adapting CRM to specific domain needs. Authors were invited to send their contribution and results on the following (and related) topics:

- CRM extensions for special uses
- Using the CRM in specific domains or subdomains
- Mapping existing metadata schemas to the CRM
- Mapping repositories and tools
- CRM and other documentation standards
- Using CRM for gazetteers and thesauri
- Using CRM in Linked Data
- CRM and Natural Language Processing
- Formalization of CRM
- Querying, searching and faceted browsing of CRM repositories
- Reasoning with CRM

The organization of the papers and the reviewing process was simplified by the tools provided by EasyChair.
I take this opportunity to thank the organizing committee for organizing such a successful event and the program committee members for their invaluable contribution, providing helpful, informative and timely reviews for all the submissions. My deepest thanks go to all the people who contributed and attended the workshop, without their submission and participation this workshop would not have been possible.
Finally, I gratefully acknowledge the ARIADNE project (funded by the European Commission under the Community's Seventh Framework Programme, FP7-INFRASTRUCTURES-2012-1-313193) for the funding sources that made the organization of the workshop possible.

Paola Ronzino (Editor)

Organizing Committee

Franco Niccolucci (PIN, Italy). Workshop Chair
Martin Doerr (FORTH, Greece)
Sorin Hermon (The Cyprus Institute, Cyprus)

Program Committee

Chryssoula Bekiari (FORTH, Greece)
Ceri Binding (University of South Wales, UK)
Paul Cripps (University of South Wales, UK)
Øyvind Eide (University of Passau, DE)
Achille Felicetti (PIN, Italy)
Reinhard Foertsch (DAI, DE)
Gerald Hiebel (University of Southern California, CA/University of Innsbruck, Austria)
Carlo Meghini (CNR-Nemis, Italy)
Dominic Oldman (British Museum, UK)
Christian-Emil Ore (University of Oslo, Norway)
Paola Ronzino (PIN, Italy)
Maria Theodoridou (FORTH, Greece)

Introduction

The CIDOC CRM ontology is an international standard currently widely accepted and adopted by different research communities and digital infrastructures to manage heterogeneous documentation (ARIADNE, PARTHENOS, 3D COFORM, 3D ICONS, iMARINE, to cite a few)[1]. It fosters interoperability among different data structures by providing the semantic definitions needed to transform different and confined information sources into a coherent and global resource, and offering a flexible system that does not impose the use of a unique standard. The CRM is used to describe the documentation process and to express the implicit and explicit concepts and relationships typically assumed in the cultural heritage documentation. By providing a common and extensible semantic framework, to which any cultural heritage information can be mapped, it prevents semantic information loss, a phenomenon that usually occurs when integrating heterogeneous resources.

Although the CIDOC CRM ontology proposes high-level concepts, which was, together with the abstractness of its concepts, one of the criticisms addressed to the ontology until recent[2], it offers the possibility to create extensions at any degree of detail, necessary to capture the full richness of the cultural heritage datasets. On the other hand, the core CRM is the common framework on which domain-specific specializations rely enabling non-domain interoperability.

Recently various extensions have been released, focusing on geographical concepts, digital provenance preservation, scientific applications and reasoning, archaeology and built structures documentation. Moreover, the CIDOC CRM ontology has proved to be fundamental to Natural Language Processing (NLP) of heritage datasets and grey literature, gazetteers and thesauri, Linked Open Data and other methods for the use and re-use of datasets and collections of digital humanities, historical and archaeological resources.
Based on these considerations, the aim of the workshop was to start a constructive debate with participants and to collect insights, issues and suggestions. This activity provided a deeper understanding on users' requirements, which were taken into account and reported to the Special Interest Group of CIDOC CRM and the ARIADNE Metadata and Standards SIG, in which the workshop orga-

[1] ARIADNE: `http://ariadne-infrastructure.eu`
PARTHENOS: `http://www.parthenos-project.eu`
3D COFORM: `http://www.3d-coform.eu`
3D ICONS: `http://www.3dicons-project.eu`
iMARINE: `http://www.i-marine.eu`

[2] Nussbaumer, Philipp and Haslhofer, Bernhard, 2007. *CIDOC CRM in Action - Experiences and Challenges*. In Research and Advanced Technology for Digital Libraries: Proceedings of the 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007. Springer Berlin Heidelberg. URL `http://dx.doi.org/10.1007/978-3-540-74851-9_61`

nizers are fully involved, providing relevant inputs and contributing in making recommendations.

During the workshop participants shared their approaches describing and showcasing systems using the CRM, exchanging experience about its practical uses and describing difficulties in its application.

This proceeding includes six selected papers presented at the workshop "Extending, Mapping, and Focusing the CRM". Every paper received two reviews, provided by the program committee members. Most of the submitted works discussed about practical applications of the CIDOC CRM global ontology for the implementation of case studies directly connected with the ARIADNE project activities.
The reason stays in that the CIDOC CRM ontology has been chosen by the ARIADNE project as the global ontology to which the archaeological datasets and collections, made available by the partner institutions, were mapped. This activity allowed the identification of common concepts and relations, which were fundamental for the implementation of the archaeological extension of the ontology, the CRMarchaeo. These mapping activities were performed using the 3M tool developed by FORTH, which provides users with a powerful graphical interface that overcomes the complexity of the global model and allows using the CRM extensions. It also acts as guide to advice users in the mapping process. The X3ML data exchange framework tool is presented in the paper *"X3ML Framework: An effective suite for supporting data mappings"*, authored by Nikos Minadakis, Yannis Marketakis, Haridimos Kondylakis, Giorgos Flouris, Maria Theodoridou, Martin Doerr, and Gerald de Jong.

Furthermore, CIDOC CRM has been chosen as the backbone ontology for the integration of heterogeneous datasets at the level of single records.
A description of this activity is reported in the paper *"Integrating heterogeneous coin datasets in the context of archaeological research"* (by Achille Felicetti, Philipp Gerth, Carlo Meghini, and Maria Theodoridou), which demonstrates the item-level integration process of archaeological archives through the use of semantic technologies.
For the implementation of this case study, a sub set of ancient coin records, provided by several European archaeological institutions, was selected. The subset thus created, was analysed to identify similar concepts and common metadata elements to enable their integration. CIDOC CRM was chosen as the conceptual model for encoding the identified entities, while some important numismatic vocabularies have been employed to improve standardisation.

Another ARIADNE-related work, *"Integrating terminological tools and semantic archaeological information: the ICCD RA Schema and Thesaurus"* (by Achille Felicetti, Ilenia Galluccio, Cinzia Luddi, Maria Letizia Mancinelli, Tiziana Scarselli, and Antonio Davide Madonna), describes the process of mapping, translation and publication in SKOS format of the RA Thesaurus. The RA Thesaurus, developed by the Italian Ministry of Cultural Heritage (MiBACT),

provides a unified and meaningful terminology for the description of archaeological objects according to the MiBACT official cataloguing standards. A detailed description of the thesaurus, is provided within the paper, together with the technologies used for the publication of the thesaurus on the web.

The paper *"Dati.CulturaItalia: a use case of publishing Linked Open Data based on CIDOC-CRM"* (by Sara Di Giorgio, Achille Felicetti, Patrizia Martini and Emilia Masci) describes the pilot project dati.culturaitalia.it, aimed at building a Linked Open Data (LOD) Service that would make open datasets from the web-portal CulturaItalia, available. The CIDOC CRM ontology was used in this case study, to transform and represent cultural heritage data. The RDF triples mapped to the CRM Erlangen were enriched with links to URIs identifying instances of internationally established RDF resources for geographic names, and instances of authority files for personal and corporate names.

The work by Achille Felicetti, Francesca Murano, Paola Ronzino, and Franco Niccolucci, *"CIDOC CRM and Epigraphy: a hermeneutic challenge"*, proposes an extension of the CIDOC CRM to encode epigraphic concepts and to model the scientific process of investigation in this domain. After identifying the main concepts involved in the study of epigraphy, and analysing the existing CIDOC CRM entities, together with those provided by the CRMsci and CRMarchaeo extensions, the authors propose to introduce the CRMepi extension. With the new classes and properties developed ad hoc, CRMepi aims at contributing to the specific needs of epigraphic documentation.

A methodological contribution to temporal knowledge is provided by the paper *"Temporal Primitives, an Alternative to Allen Operators"* (by Manos Papadakis, and Martin Doerr). The paper discusses the limits of the Allen Interval Algebra set of operators, which fails in observation-driven fields like stratigraphy. In such cases, incomplete temporal information yields a disjunctive set of Allen operators, which affects RDF reasoning since it leads to expensive queries containing unions. To address this deficiency, the authors introduce a set of basic temporal primitives which are employed in an extension of CIDOC CRM. The flexible representation proposed by the authors can describe any Allen operator as well as scenarios with further temporal generalization using conjunctions of primitives. An extension to the basic set of primitives is also proposed, introducing fuzzy primitives that can model temporal topologies with imprecise boundaries that generalize over precise boundary models.

Paola Ronzino

PIN, Prato, Italy

X3ML Framework: an Effective Suite for Supporting Data Mappings

Nikos Minadakis[1], Yannis Marketakis[1], Haridimos Kondylakis[1], Giorgos Flouris[1], Maria Theodoridou[1], Gerald de Jong[2], and Martin Doerr[1]

[1]Institute of Computer Science, FORTH-ICS, Greece
[2]Delving B.V. The Netherlands
{minadakn,marketak,kondylak,fgeo,maria,martin}@ics.forth.gr
gerald@delving.eu

Abstract. The aggregation of heterogeneous data from different institutions in cultural heritage and e-science has the potential to create rich data resources useful for a range of different purposes, from research to education and public interests. In this paper, we present the architecture and functionality of X3ML data exchange framework, that handles effectively and efficiently the schema mapping, URI definition and generation, and data transformation steps of the provision and aggregation process. The X3ML framework is based on the *X3ML mapping definition language* that offers the building blocks for describing both schema mappings and URI generation policies, and the *X3ML engine*, that handles the URI generation and the data transformation. The X3ML framework supports the cognitive process of mapping and it has a lot of advantages compared to other existing tools including that the schema mappings are expressed in a declarative way, and are both human and machine readable allowing domain experts to understand them, the schema matching and the URI generation policies comprise different distinct steps in the exchange workflow, and follow different life cycles. Furthermore X3ML is symmetric and potentially invertible allowing bidirectional interaction between providers and aggregator and thus supporting not only a rich aggregators' repository but also corrections and improvements in the providers' data bases.

Keywords: Data Mappings, Data Aggregation, URI Generation

1 Introduction

Managing heterogeneous data is a challenge for cultural heritage institutions, such as archives, libraries, and museums, but equally for research institutes of descriptive sciences such as geology, biodiversity, clinical studies etc. These institutions host and develop various collections with heterogeneous material, often described by different metadata schemas. In order to provide uniform access to heterogeneous and autonomous data sources, complex query and integration mechanisms have to be designed and implemented.

In order to allow data transformation and aggregation, it is required to produce mappings, to relate equivalent concepts or relationships from the source schemata to the aggregation schema, i.e. the target schema, in a way that facts

described in terms of the source schema can automatically be translated into descriptions in terms of the target schema, or "enterprise model" as Calvanese et al. [6] describe it. This is the mapping definition process and the output of this task is the mapping, i.e., a collection of mapping rules.

In this paper we describe the X3ML framework, which is able to support the data aggregation process by providing mechanisms of data transformation and URI generation. Mappings are specified with the X3ML mapping definition language which is a declarative, human readable language that supports the cognitive process of a mapping. Unlike XSLT, that can only be understood by IT technicians, X3ML can be understood by non-technical people, so a domain expert is capable of testing the semantics, reading and validating the schema matching. This model carefully distinguishes between mapping information from the domain experts who know and provide the data and that created by the IT technicians who actually implement data translation and integration solutions, and serves as an interface between both.

A common problem of a schema matching and transformation process is that the IT experts do not fully understand the semantics of the schema matching and the domain experts do not understand how to use the technical solutions. For this reason, in our approach the URI generation and the schema matching processes are separated, so the schema matching can be fully performed by the domain expert and the URI generation by the IT expert, and therefore solving the bottleneck that requires that the IT expert understands the mapping. Furthermore this keeps the schema mappings between different systems harmonized since the schema mappings definitions do not change in contrast to the URIs that may change between different institutions and are independent of the semantics. XSLT and R2RML have tightly coupled the URI generation from the schema matching processes.

Our approach completely separates the definition of the schema matching from the actual execution. This is important because different processes might have different life-cycles; in particular the schema matching definition has a different life-cycle compared to the URI generation process. The former is subject to more sparse changes compared to the latter.

The remainder of this paper is organized as follows: Section 2 discusses the related work. Section 3 discusses about the background. Section 4 describes the details of the X3ML framework. Section 5 enumerates different usages of the framework and provides the evaluation results. Finally Section 6 concludes and discusses about the future directions of our work.

2 Related Work

Mapping relational databases (RDB) to RDF became a quite active field the last few years. This happens as the majority of data currently published on the web are still stored in relational databases with local schemata and local identifiers. There are several solutions that fall into this category. Direct Mapping [13] maps automatically relational tables to RDF classes and attributes to RDF properties. D2R MAP [5] is a declarative language to describe mappings between relational

databases and OWL/RDFS ontologies. Triplify [2] maps HTTP-URI requests onto RDB queries and translates the resulting relations into RDF statements. R2RML[1] which is a mapping language proposed by W3C in order to standardize RDB to RDF mappings.

In addition there are tools, that provide mappings from XML to RDF leading to mappings in the syntactic level rather than in the semantic level. Tools in this category include tools based on XSLT (i.e. Krextor [10], AstroGrid-D[2]), tools based on XPATH (i.e. Tripliser[3]) and XQUERY (i.e. XSPARQL [4]). There are also other approaches that exploit mapping technologies to publish their data as linked data. For example the Smithsonian American Art Museum used KARMA [17] to publish their data as linked data, a tool trying to automate the mapping process allowing users to adjust the generated mappings. However, there is still no clear distinction on the work of the domain and the IT experts which perplexes the whole workflow. KARMA uses R2RML model so it inherits the issue of tight coupling between the schema matching and the URI generation.

Finally there are similar works that map CSV files to RDF. XLWrap's mapping language [11] provides conversions from CSV and spreadsheets to RDF data model. Mapping Master's M2 [14] converts data from spreadsheets into OWL statements.

All these different approaches prove that there is no standard model to support mapping of data sources other than relational, the technologies used are too complex to be used by the domain experts and the whole workflow is not well-defined. Compared to these works our work (a) uses a simple model for defining the mappings in a way that is comprehensible and readable from the domain experts, (b) is generic because the mapping definitions are not tied to the implementation of the data transformation engine, (c) supports incremental changes of source and target schema, (d) supports customized URI generation policies and (e) promotes the collaborative work of experts with different roles on the mapping process.

3 Background

The main pillar of our work is the Synergy Reference Model (for short SRM) which is an initiative of the CIDOC CRM Special Interest Group[4]. It is a reference model for a better practice of data provisioning and aggregation processes, primarily in the cultural heritage sector, but also for e-science. It is based on experience and evaluation of national and international information integration projects. It defines a consistent set of business processes, user roles, generic software components and open interfaces that form a harmonious whole. Currently a draft version of the model is available online[5], still being evolved and enriched. The goal of SRM is to: (a) describe the provision of data between providers

[1] http://www.w3.org/TR/r2rml/
[2] http://www.gac-grid.de/project-products/Software/XML2RDF.html
[3] http://daverog.github.io/tripliser/
[4] http://www.cidoc-crm.org/who_we_are.html
[5] http://www.cidoc-crm.org/docs/SRM_v1.4.pdf

and aggregators including associated data mapping components, (b) address the lack of functionality in current models (i.e. OAIS [12]) and practice, (c) incorporate the necessary knowledge and input needed from providers to create quality sustainable aggregations and, (d) define a modular architecture that can be developed and optimized by different developers with minimal inter-dependencies and without hindering integrated UI development for the different user roles involved.

SRM aims at identifying, supporting or managing the processes needed to be executed or maintained between a provider (the source) and an aggregator (the target) institution. It supports the management of data between source and target models and the delivery of transformed data at defined times, including updates. This includes a mapping definition, i.e., specification of the parameters for the data transformation process, such that complete sets of data records can automatically be transformed. A graphical representation of the data provisioning workflow is shown in Fig. 1.

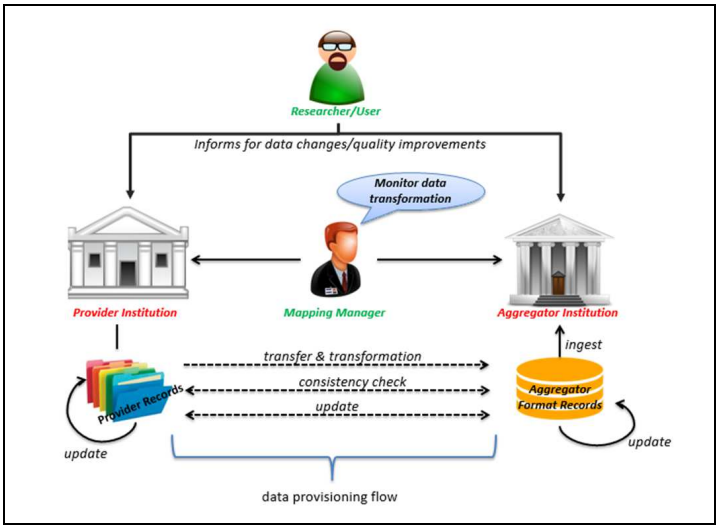


Fig. 1. The data provisioning workflow

The main steps of the data provisioning workflow are:

- **Schema matching**: source and target schema experts (a.k.a the domain experts) define a schema matching which is documented in a schema matching definition file. This file should be human and machine readable and it is the ultimate communication mean on the semantic correctness of the mapping.
- **Instance generation specification**: in this step the URI generation and datatype conversion policies are defined for each instance of a target schema class referred to in the matching. In this step only IT experts are involved and domain experts have no interest or knowledge about it.
- **Terminology mapping**: the terminology mappings between source and target data/terms are defined. Providers may use anything from intuitive lists of uncontrolled terms up to highly structured third party thesauri.

- **Transformation**: once the mapping definition has been finalized (and all syntax errors are resolved) the data needs to be transformed, producing a set valid target records. The transformation process itself may run completely automatically. In the case where any issues arise, the aggregator can resolve them on a temporary or permanent basis but it is also possible that these records are sent back to the provider for further analysis and resolution.
- **Ingestion**: once records are transformed, an automated translation for source terms using a terminology map follows. The transformed records will then, be ingested into the target system.
- **Change detection**: after the ingestion of the records all changes that may affect the consistency of provider and aggregator data are monitored. SRM describes 18-20 different updating and transformation reasons and is the only framework at the moment which takes the maintance into account.

4 The X3ML framework

The X3ML framework comprises the X3ML Mapping Definition Language and the X3ML Engine. Below we will describe them.

4.1 X3ML Mapping Definition Language

The X3ML mapping definition language is an XML based language which describes schema mappings in such a way that they can be collaboratively created and discussed by experts. The X3ML language was designed on the basis of work that started in FORTH in 2006 [9] and emphasizes on establishing a standardized mapping description which lends itself to collaboration and the building of a mapping memory to accumulate knowledge and experience. It was adapted primarily to be compliant with the DRY principle (avoiding repetition) and to be more explicit in its contract with the URI Generating process. X3ML separates schema mapping from the concern of generating proper URIs so that different expertise can be applied to these two very different responsibilities.

Schema matching: Schema matching is performed by domain experts who need to be concerned only with the correct interpretation of the source schema. The structure of X3ML is quite easy to understand consisting of: (a) **a header** that contains basic information (title, description, contact persons), the source and target schemata and sample record, and (b) **a series of mappings** each containing a domain (the main entity that is being mapped) and a number of links which consist of a path and a range. Each link describes the relation (path) of the domain entity to the corresponding range entity.

The basic mapping scheme and the corresponding XML structure is shown in Fig. 2. Each *entity-relation-entity* of the source schema is mapped individually to the target schema and can be seen as self-explanatory, context independent proposition. An X3ML structure consists of:

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

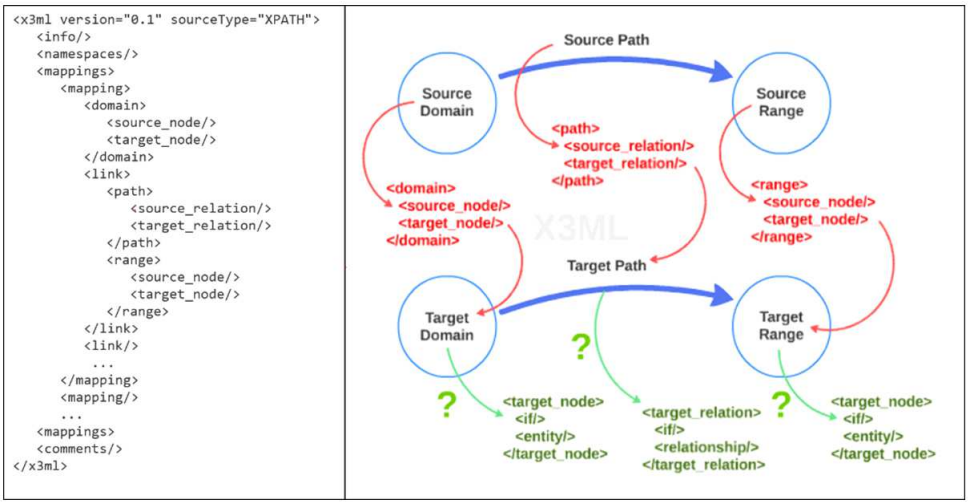


Fig. 2. The structure of an X3ML mapping

- the proper target path
- the mapping between source path and target path

The X3ML mapping definition language supports 1:N mappings and uses the following special constructs:

- **intermediate nodes** used to represent the mapping of a simple source path to a complex target path (a sequence of path-{entity-path}).
- **constant expression nodes** used to assign constant attributes (e.g. a constant type) to an entity.
- **conditional statements** within the target node and target relation support checks for existence and equality of values and can be combined into boolean expressions.
- **"Same as" variable** used to identify a specific node instance for a given input record that is generated once but is used in a number of locations in the mapping.
- **Join operator (==)** used in the source path to denote relational database joins
- **info and comment blocks** throughout the mapping specification bridge the gap between human author and machine executor.

The tools that are currently used to produce the X3ML mapping definition are restricted to consuming XML input records[6]. As a result, XPath is used to specify the source elements and paths which are evaluated within the context of the source domain. There is ongoing work for an extended version that will also support RDF input (see Section 4.5).

URI generation policy: The definition of the URI generation policy follows the schema matching and is performed usually by an IT expert who must ensure that the generated URIs match certain criteria such as consistency and

[6] `http://www.ics.forth.gr/isl/3M/`

uniqueness. A set of predefined URI generators (UUIDs, literals) and templates are available but any URI generating function can be implemented and incorporated in the system. In the X3ML definition, the target domain and range contain the functions that generate URIs or literals.

The result of the schema matching and URI generation policy steps is a complete X3ML mapping definition file that will be fed to the X3ML engine for the transformation of the data.

Fig. 3 shows how a simple relational database entry that specifies the weight of a coin is mapped and expressed with respect to the CIDOC CRM schema[7]. The XML structure for the mappings of this example can be found online[7].

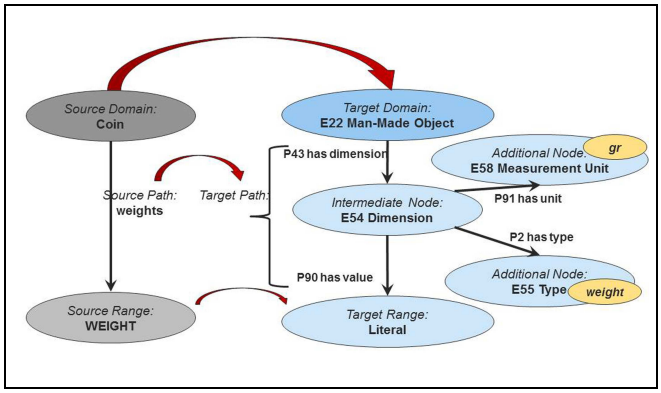


Fig. 3. Mapping relational db data to CIDOC CRM

4.2 X3ML Engine

The X3ML engine realizes the transformation of the source records to the target format. The engine takes as input the source data (currently in the form of an XML document), the description of the mappings in the X3ML mapping definition file and the URI generation policy file and is responsible for transforming the XML document into a valid RDF document which is equivalent with the XML input, with respect to the given mappings and policy. The engine has been originally implemented in the context of the CultureBrokers project co-funded by the Swedish Arts Council and the British Museum.

4.3 Design, Architecture and Implementation

The X3ML Engine has been designed with respect to the following design principles:

– *Simplicity.* It is easier to create complicated things than it is to find the simplicity in something that would otherwise be complex. One important way to achieve simplicity and clarity is by carefully naming things so that their meaning is as obvious as possible to the naked eye.

[7] `http://139.91.183.44/x3mlEditor/ViewPublished?type=Mapping&id=1`

- *Transparency.* The most important feature of X3ML is its general application to mapping creation and execution and hopefully its longevity. People must be able to easily understand how it works. The cleaner the core design of the engine and X3ML language, and the clearer its documentation, the more readily it will get traction and become the basis for future mappings.
- *Re-use of Standards and Technologies.* The best way to build a new software module is to carefully choose its dependencies, and keeping them as small as possible. Building on top of proven technologies is the quickest way to a dependable result.
- *Facilitating Instance Matching.* This involves extracting semantic information with the intent of generating correct instance URIs.

Fig. 4 depicts the main components of the engine. The *Input Reader* component is responsible for reading the input data (currently we support XML documents, however as we describe later in Section 4.5 more formats will be supported using proper extensions). The *X3ML Parser* component is responsible for reading and manipulating the X3ML mapping definitions. The component *RDF Writer* outputs the transformed data into RDF format. The *Instance Generator* component produces the URIs and the labels based on the descriptions that exist in the mappings and finally the *Controller* component coordinates the entire process.

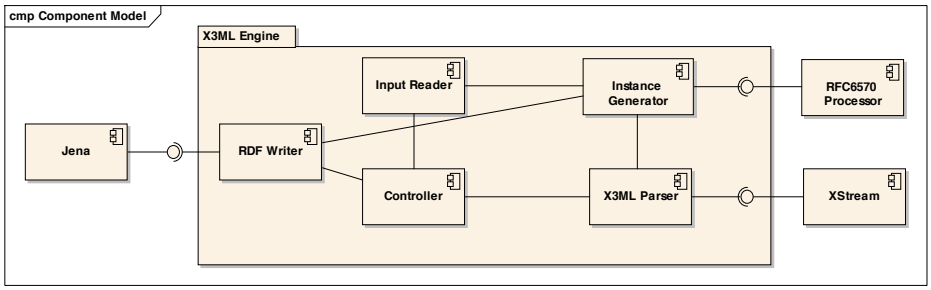


Fig. 4. The main components of X3ML Engine

The X3ML engine has been implemented in Java, producing a single artifact in the form of a JAR file which contains the engine software. For supporting the functionality of the main components we exploited a set of third-party software libraries. For instance we used XStream[8] for parsing XML-based documents, Handy URI Templates[9] to support the generation of valid URIs and Jena[10] for building the RDF output. The source code of the X3ML engine framework is available under the Apache license and can be found at `https://github.com/delving/x3ml`.

[8] `http://x-stream.github.io/`
[9] `https://github.com/damnhandy/Handy-URI-Templates`
[10] `https://jena.apache.org/`

4.4 Functionality

The X3ML engine takes as input source XML records and generates RDF triples consisting of subject, predicate, and object. The subject and the object are "values", generally consisting of URIs, but objects can also be labels or literal values.

The generation of values (URIs, or literals) is being handled by the Instance Generator component. The following block shows two configurations; for generating (a) URIs and (b) label values.

```
<instance_generator name="[gen-name]">
    <arg name="[arg-name]" type="[arg-type]">[arg-value]</arg>
    ...
</instance_generator>
<label_generator name="[gen-name]">
    <arg name="[arg-name]" type="[arg-type]">[arg-value]</arg>
    <arg name="language" type="constant">[language-code]</arg>
    ...
</label_generator>
```

For each entity there must exist one `instance_generator` and any number of subsequent `label_generator` blocks. The argument `type` allows for choosing between `xpath` and `constant` and there is a special argument type called `position` which gives the value generator access to the index position of the source node within its context. The argument with the name `language` defines the language tag of the generated value. If it is empty then it is implied that the generated value will not have it (i.e. in the case of number values). The engine provides default implementations for producing: (a) URIs, (b) UUIDs, (c) literal values and (d) constant values.

The Instance Generator component is configured through an XML file (which is given as input in the X3ML engine). When URIs are to be generated on the basis of source record content, it is wise to leverage existing standards and reuse the associated implementations. For template-based URI generation there is available the RFC 6570 [8] standard. So, the component uses an existing implementation library as described in Section 4.3. Whenever the required URIs or labels cannot be generated by the default generators, the simple templates, or the URI templates, it is always possible to insert a special generator in the form of a class implementing the *InstanceGenerator* component interfaces.

4.5 Configuration/Extensibility

As already discussed the current version of the X3ML engine, takes as input the source data in the form of an XML document. One extension (which is currently under development) is to support other types of input. To this direction we have started working on supporting RDF input. This requires several modifications in the design and implementation of the engine. More importantly the basic construct that we use for reading the source data will be an RDF model (i.e. Jena, Sesame), so instead of XPATH we will be able to use SPARQL [16]. Furthermore we will enhance the Instance Generator component since we will be able to carry the URIs from the source data to the target data if needed.

One apparent advantage of this approach is that the framework will support input and output of the same format. This sparked the light to investigate another direction; that of invertible X3ML mappings. In an invertible X3ML mapping, one can identify, in a unique manner, (and consequently regenerate) the data in the source dataset that led to the creation of each piece of data in the target dataset. Based on this idea, below we formalize the notion of invertibility, by trying to identify how X3ML maps the source data to the target data.

In particular, we view an X3ML mapping as an association between a "pattern" (say P_s) in the source dataset with a "pattern" (say P_t) in the target dataset. This association essentially describes what to put in the target dataset (P_t) whenever P_s is encountered in the source dataset. Formally, we model P_s and P_t as SPARQL graph patterns [15, 1] so an X3ML mapping m is just a pair (P_s,P_t) of SPARQL graph patterns.

Then, given a set of X3ML mappings (say M), we say that M is invertible if and only if we can guarantee that whenever a pattern (say P_t) is found in the target dataset, we can identify in a unique manner the pattern P_s that generated it (i.e., caused its inclusion in the dataset). To determine that, we look at each P_t in M (and its corresponding P_s), and identify those mappings that can potentially lead to the same triples to be generated from different source triples.

5 Evaluation and Usage

The X3ML engine is being exploited by several European projects. Specifically, the ARIADNE project[11] initiated several mapping activities using X3ML engine, to convert existing schemata of archaeological data to CIDOC CRM and its extension suite. The partners of ARIADNE project had extensively used X3ML for the definition of mappings from various categories of databases, including archeological museums, buildings, ancient Roman coins, and more. The ResearchSpace project[12] is developing a collaborative environment for humanities and cultural heritage research. The project has been using X3ML for the mapping and transformation of the Rijksmuseum, the British Museum and the Yale Center for British Art (YCBA) data. Specifically for the case of the Rijksumuseum domain experts from both Rijksumuseum and the British Museaum were able to succesfully map and transform their data without the assistance of any IT expert. X3ML engine is also being exploited by the transformation services of the Greek national implementation of the European LifeWatch [3] infrastructure for biodiversity to transform biodiversity metadata/data such as Darwin Core formats to a CIDOC CRM family semantic models.

To evaluate[13] the performance of the X3ML engine we used an XML input and a X3ML mapping example coming from the ARIADNE Project as a base to

[11] `http://www.ariadne-infrastructure.eu/`

[12] `http://www.researchspace.org/`

[13] The experiments were carried out on a PC with an Intel i7 processor, 8GB RAM, running Windows 7 32 bit.

produce synthetic data that was provided as input to the X3ML engine. Three X3ML mapping files were created containing 10,100 and 1000 mappings and 4 XML input files containing 10,100,1000 and 10000 records. Fig. 5 displays the evaluation results. We can observe that the overall time depends on both the number of mappings and the size of the input. For example, as we can see from the evaluation results, the time required for data transformation is approximately one second when the size of the input is low (10 records) even if the mappings are many(from 10 to 1000). As the size of the input increases however, the overall time that is required increases as well. Note, that the total number of output records is the total number of input records multiplied with the number of mappings (i.e. 10 input records with 10 mappings will produce 100 output records). Concluding, we can see that the execution time is affected equally by the number of the mappings and the records, and it is related with the number of the links that are created during the transformation process.

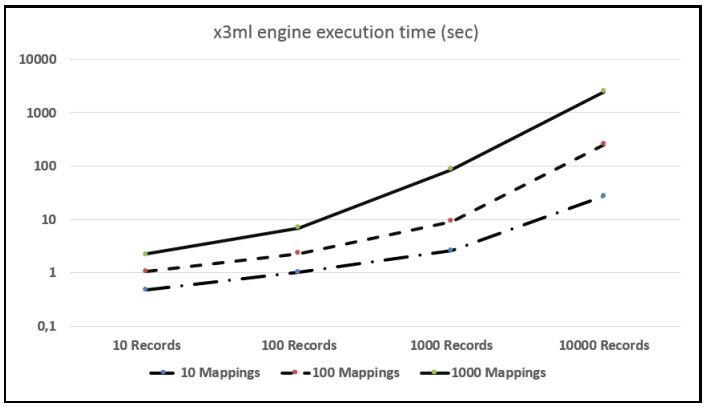


Fig. 5. x3ml Engine Evaluation Results

6 Conclusion and Future Work

This paper presents a novel framework for the management of the core processes needed to create, maintain and manage mapping relationships between different data sources. We described the X3ML mapping definition language that offers the building blocks for describing both schema mappings and URI generation policies and X3ML engine, a tool that supports the transformation process and the generation of URIs and values and is characterized by its scalability in terms of number of providers, consistent mappings and related end up processes. We demonstrated some of our experiences on using the aforementioned framework and discuss about the evaluation results. In future we plan to continue working on the extended version of the framework that will support different types on input (i.e. RDF documents) and investigate the invertible X3ML mappings functionality.

Acknowledgement

This work was partially supported by the project *PARTHENOS* (H2020 Research Infrastructures, 2015-2019), the project *ARIADNE* (FP7 Research Infrastructures, 2013-2017), and the *LifeWatch Greece* project (National Strategic Reference Framework, 2012-2015).

References

1. M. Arenas, C. Gutierrez, and J. Pérez. *On the Semantics of SPARQL*. Springer, 2010.
2. S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: light-weight linked data publication from relational databases. In *Proceedings of the 18th international conference on World wide web*, pages 621–630. ACM, 2009.
3. A. Basset and W. Los. Biodiversity e-science: Lifewatch, the european infrastructure on biodiversity and ecosystem research. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 146(4):780–782, 2012.
4. S. Bischof, S. Decker, T. Krennwallner, N. Lopes, and A. Polleres. Mapping between rdf and xml with xsparql. *Journal on Data Semantics*, 1(3):147–185, 2012.
5. C. Bizer. D2r map-a database to rdf mapping language. 2003.
6. D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Description logic framework for information integration. In *KR*, pages 2–13, 1998.
7. M. Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.
8. J. Gregorio, R. Fielding, M. Hadley, M. Nottingham, and D. Orchard. Rfc 6570: Uri template. *Internet Engineering Task Force (IETF) Request for Comments*, 2012.
9. D. P. Haridimos Kondylakis, Martin Doerr. Mapping language for information integration. *Technical Report ICS-FORTH*, 385, 2006.
10. C. Lange. Krextor-an extensible framework for contributing content math to the web of data. In *Intelligent Computer Mathematics*, pages 304–306. Springer, 2011.
11. A. Langegger and W. Wöß. *XLWrap–querying and integrating arbitrary spreadsheets with SPARQL*. Springer, 2009.
12. B. Lavoie. Meeting the challenges of digital preservation: The oais reference model. *OCLC Newsletter*, 243:26–30, 2000.
13. T. B. Lee. Relational databases on the semantic web. *Design Issues (published on the Web)*, 1998.
14. M. J. O'Connor, C. Halaschek-Wiener, and M. A. Musen. Mapping master: A flexible approach for mapping spreadsheets to owl. In *The Semantic Web–ISWC 2010*, pages 194–208. Springer, 2010.
15. J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of sparql. In *International semantic web conference*, volume 4273, pages 30–43. Springer, 2006.
16. E. Prud'Hommeaux, A. Seaborne, et al. Sparql query language for rdf. *W3C recommendation*, 15, 2008.
17. P. Szekely, C. A. Knoblock, F. Yang, X. Zhu, E. E. Fink, R. Allen, and G. Goodlander. Connecting the smithsonian american art museum to the linked data cloud. In *The Semantic Web: Semantics and Big Data*, pages 593–607. Springer, 2013.

Integrating Heterogeneous Coin Datasets in the Context of Archaeological Research

Achille Felicetti[1], Philipp Gerth[3], Carlo Meghini[4], and Maria Theodoridou[2]

[1] PIN, VAST-LAB, Prato, Italy
[2] FORTH-ICS, Greece
[3] DAI, Germany
[4] CNR-ISTI, Italy
achille.felicetti@pin.unifi.it,carlo.meghini@cnr.it
maria@ics.forth.gr,
philipp.gerth@dainst.de,
http://www.ariadne-infrastructure.eu/

Abstract. This paper describes the activities carried out under the ARIADNE project to demonstrate the item-level integration process of archaeological archives through the use of semantic technologies. To this end, some ancient coin records, coming from the archives of important European archaeological institutions, were selected. The subset thus created, has been carefully analysed by means of specific tools to identify similar concepts and common metadata elements that could serve as the basis for integration. CIDOC CRM was chosen as the conceptual model for encoding the identified entities, while some important numismatic vocabularies have been employed to improve standardisation. The implementation phase has benefited from the use of advanced tools for mapping and conversion of the original information in a semantic form (RDF), the creation of a triple store to place the newly integrated data and the necessary interfaces for accessing and querying them.

Keywords: Coins, Integration, CIDOC CRM, RDF

1 Introduction

Information Technology (IT for short) is quickly and widely conquering the Humanities: more and more scholars use IT methods and tools to build, access, share and preserve the knowledge that they generate in their daily research activities. This phenomenon also concerns the past: there are many projects that aim at reviving datasets and collections that have been generated in previous endeavours, whether they are in analog or digital form, in order to treat these data with the novel IT capabilities. The motivation behind this vast movement is rather obvious to anybody minimally familiar with research: the quality and quantity of knowledge generated by research activities is positively correlated with the amount of information and knowledge used in the process: the larger the latter, the greater the former.

The IT world is responding to the demand so generated by the Humanities in a positive way; new methods and tools are constantly produced that facilitate the work of the humanist, in the context of projects where IT specialists and scholars actively collaborate to the fullfillments of the project objectives. Lately, this collaboration is taking place through research infrastructures, perhaps the most relevant contribution of IT to the scientific world. This paper exemplifies one such collaboration, taking place in the context of ARIADNE[5], an FP7-INFRASTRUCTURES-2012-1 EU project (Grant agreement no: 313193), aiming at building a research infrastructure in the field of Archaeology. Specifically, the work described in this paper focuses on the integration of heterogeneous datasets containing information about coins.

From an IT point of view, this work is classified as a *data integration* activity, taking as input several datasets whose contents overlap in time, space and subject, and producing as output a novel dataset. The resulting novel dataset contains all the information of the original datasets, but integrated in a coherent whole that can be queried to discover knowledge previously inaccessible. In order to obtain the integrated dataset, several important and non-trivial problems had to be solved.

1. First of all, a deep, accurate and extensive analysis of the input data has been carried out, in order to determine the information space of each dataset, both in terms of the attributes covered and of the values used to instantiate such attributes. The analysis has allowed us to obtain important syntactic and semantic information, addressing the conceptual and lexical space of each dataset. More importantly, it has confirmed the validity of the project, by showing an effective overlapping of the knowledge in the given datasets.
2. Once the individual information space of each dataset was understood, the design of the integration has started, aimed at devising a common ontology that could serve as conceptual backbone of the integrated dataset. In achieving this specific objective, we have built on the experience gained in previous projects, and avoided to invent yet another ontology. Rather, we have relied on the CIDOC CRM ontolgoy [CRM15], the ISO 21127:2006 standard that is being successfully employed for documentation and data integration in the domain of cultural heritage since few decades. We have therefore ascertained that the CRM was rich enough to cover the integrated information space, and have set out to identify the vocabularies that we could use for integrating the values used in each dataset. This stage of the project proved difficult enough, due to the lack of individual vocabularies that could cover the value information space with the required generality and exhaustivity. A detailed account of this activity is given in the paper.
3. We have then entered the implementation stage of our project, aimed at devising mappings linking the attributes of the input datasets to the properties of the CRM, and the attribute values of the datasets to the vocabularies chosen for the integrated dataset. Also for this stage of the project we have

[5] `http://www.ariadne-infrastructure.eu`

relied on an existing tool, the X3ML suite [MMK+15], including an editor (3M[6]) for the mapping specification and an engine (X3ML engine[7]) for the mapping execution. The X3ML suite fills several blocks of a general architecture for data integration, named SYNERGY [ODdJ+14], [DFdJ+15], an initiative of the CIDOC CRM Special Interest Group[8], currently employed in several ongoing projects, such as the already mentioned ARIADNE, the just started PARTHENOS[9], Lifewatch[10], ResearchSpace[11], ITN-DCH[12] and Cultural Heritage Imaging [13].

4. We are presently executing the mappings and implementing the persistence of the generated dataset. To this end, we intend to exploit Semantic Web languages and technologies for representing and implementing the integrated dataset, so as to maximize interoperability and therefore re-use.

The different parts of the paper account for the just described phases of the project, providing detailed descriptions of the problems encountered and of the methods and tools employed to solve them. For the future, we foresee two main activities:

- specification of the queries on the integrated dataset, with special care for those returning knowledge coming from at least two of the input datasets;
- implementation of an access facility to the integrated dataset, both on the web and on the ARIADNE infrastructure and possibly more.

An articulation of these activities is provided in the concluding section of the paper.

2 List and Description of the Archives to be Integrated Within ARIADNE.

Numismatics is a very traditional science with a lot of experience and early initiatives in standardization of the existing data (f.e. [BV78]). In recent years the numismatics excels in terms of Linked Open Data in the Digital Humanities with a high grade of accessible datasets and standardized vocabulary. One major collaborative project is Nomisma.org[14], supported by a lot of institutions. Nomisma.org serves as a authorative resource in the numismatics. It collects and provides URIs to common numismatic concepts and terms. Furthermore

[6] http://www.ics.forth.gr/isl/3M/
[7] https://github.com/isl/x3ml
[8] http://www.cidoc-crm.org/who_we_are.html
[9] http://www.parthenos-project.eu
[10] https://www.lifewatchgreece.eu
[11] http://www.researchspace.org/
[12] http://www.itn-dch.eu/
[13] http://culturalheritageimaging.org
[14] http://nomisma.org/

a whole ontology[15] was created, which is used to integrate the open available databases. The ontology provides an easy understandable way for numismatists to describe their dataset, but as it is just limited to the numismatics, its very domain specific, other than the generic approach of CIDOC-CRM. Overall, numismatics provides a very good starting point for the item-level integration of archaeological datasets, as it is highly standardized and data is widely available to demonstrate the usefulness of using ontologies.

2.1 The dFMRÖ archive

Digitale Fundmünzen der Römischen Zeit in Österreich (dFMRÖ, digital Coin-finds of the Roman Period in Austria) is an online MySQL database of the Numismatic Research Group of the Austrian Academy of Sciences [dFM07]. Since the 1990s it documents coin-finds from the Celtic and Roman Period that have been published in various printed volumes of the FMRÖ (Fundmünzen der Römischen Zeit in Österreich / Coin-finds of the Roman Period in Austria) from the 1970s up to 2007. Starting with a Microsoft Access database, it was set up in its current form in 2007 and hosts about 76.000 finds. All coins in the database were found in Austria from the Celtic and Roman period (actually the entire Antiquity), registered properly so no illegal finds are included and most of them already published by the various projects of the FMRÖ. Because of a former project cooperation, since 2007 it also lists coins found in Romania. These are the coins that were published in: "Colonia Ulpia Traiana Sarmizegetusa", the first volume of "Coins from Roman sites and collections of Roman coins from Romania". The coins represent an important part of the Austrian cultural heritage.

The dFMRÖ archive was chosen as the first hands-on exercise to map a relational data base schema to CIDOC CRM, since it represents a large class of well-defined traditional databases. A sample XML record from the dFMRÖ archive is shown in the Appendix.

2.2 Numismatic archives from the COINS project

Another source of information we have taken into account, comes from two numismatic archives already used within the COINS project. They include a set of 1670 numismatic records coming from the Cambridge Fitzwilliam Museum archive (FWM) and a set of 630 records coming from the Sprintendenza Archaeologica di Roma (SAR) database.

The COINS project (Combat On-line Illegal Numismatic Sales) aimed at providing a substantial contribution to the fight against illegal trade and theft of coins by using state-of-art Information Technology. The project developed standardized inventories by integrating legacy archives encoded in different formats and using different languages. The creation of a reference collection of Roman and Greek coins was also one of the most relevant outcomes of the project.

[15] `http://nomisma.org/ontology`

The FWM archive: The FWM subset comes from the Department of Coins and Medals of the Fitzwilliam Museum Database, recording information on medals and coins of different types age, discovered during excavations or coming from various acquisitions or donations, currently kept by the FW museum. Relevant fields used by FWM archive include: coin maker, production location, mint, coin type, category, coin name, inscription, dimensions, production technique, references to images. Databases also include notes concerning record creation and modification, date and time, museum acquisition information. An XML example of an FWM record is shown in the Appendix.

The SAR archive: The SAR database (originally a Microsoft Access DB) was created for the cataloguing of archaeological finds of monetary type managed by the Archaeological Superintendence of Rome, coming from public and private collections and from archaeological excavations made in the city of Rome and its immediate surroundings. The main purpose of the archive is to record to provide the date, the accurate descriptions (by indicating the precise origin or place of issue) and the physical characteristics of the various coins. In addition it also shows the conditions of discovery (excavation, auction, seizure, donation, etc.), the state of preservation and the current location location (museum, superintendences, collections and so on).

SAR database, in addition to the FWM fields reported above, also provides information concerning coins physical features and physical conditions, the region in which a specific coin was minted (apart the exact location), specific information on chronology (i.e. the age, century or period during which coin minting took place), obverse/reverse inscriptions of iconography and the current location of the specific exemplar the record refers to. An XML example of a SAR record is shown in the Appendix.

2.3 Arachne

Arachne[16] is the central object database of the German Archaelogical Institute (DAI). Currently it contains more than 2,000,000 images with corresponding metadata and over 300,000 highly structured descriptions of artifacts of archaeological interest. Also Arachne allows research projects to store, manage and publish their data in online available catalogs. Coming out of digitized museum inventory and research project data, there are currently 485 coins with varying metadata quality. Some are of excellent quality, as the 107 coins with figures related to harbours coming from the DFG founded "SPP-Häfen". Those provide beside a detailed description extensive information about bibliographic references and dating opinions of different authors.

[16] `arachne.dainst.org`

2.4 iDAI.field

Since the first usage in 2005, the field research database iDAI.field was adopted by around 35 archaeological projects. The modular system contains also a find module with specific attributes for coins, which were found during excavations or surveys. For a first integration test 517 coins of the Pergamon project were used with detailed information about the archaeological context. An XML example of selected iDAI.field attributes is shown in the Appendix.

2.5 MuseiD-Italia collections

We are also investigating the possibility to integrate the collections of MuseiD-Italia, the digital library integrated in CulturaItalia. The data are in CIDOC CRM form and can be extracted via the OAI-PMH of the repository. MuseiD-Italia includes several collections of coins from Italian museums such as:

1. *Museo archeologico nazionale di Venezia*
   Il medagliere: serie romana - imitazioni o falsificazioni moderne, 86 coins
   Il medagliere: serie greca e bizantina, 758 coins
   Il medagliere: serie romana e barbarica, 2307 coins
2. *Museo archeologico nazionale di Crotone*, Reperti archeologici e Numismatica, 31 coins
3. *Collezione Museo Archeologico Nazionale - Reggio di Calabria*, 136 coins
4. *Collezione numismatica Museo Archeologico Nazionale di Altamura*, 99 coins
5. 3008 coins from Regione Umbria

3 Mapping Operations at a Logical and Practical Level

The aggregation and integration of the datasets described in the previous section was chosen as an appropriate use case for the ARIADNE infrastructure to prove that it is possible to create a rich common repository at a data item level, useful for a range of different purposes, from research to education and engagement. In this section we present the aggregationn workflow that we have followed in order to map a set of source databases to a common target ontology and trasform the original records to resources of the common, integrated repository.

The process of aggregating a set of databases consists of four major steps as shown in figure 1:

1. **Schema matching**: this first step produces mappings from the schema of each source dataset (source schema) to the common CIDOC CRM ontology (target schema). It is very important that the mappings obtained during schema matching preserve as much as possible the meaning of the source schemas fields. To this end, a close collaboration is required amongst domain experts, who know the semantics of the source schemas, pivot ontologists, who know the semantics of the chosen pivot ontology, and the IT experts who guide the others on using the mapping specification tools; these tools include both the language for encoding the mappings and the software for creating and managing the mappings.

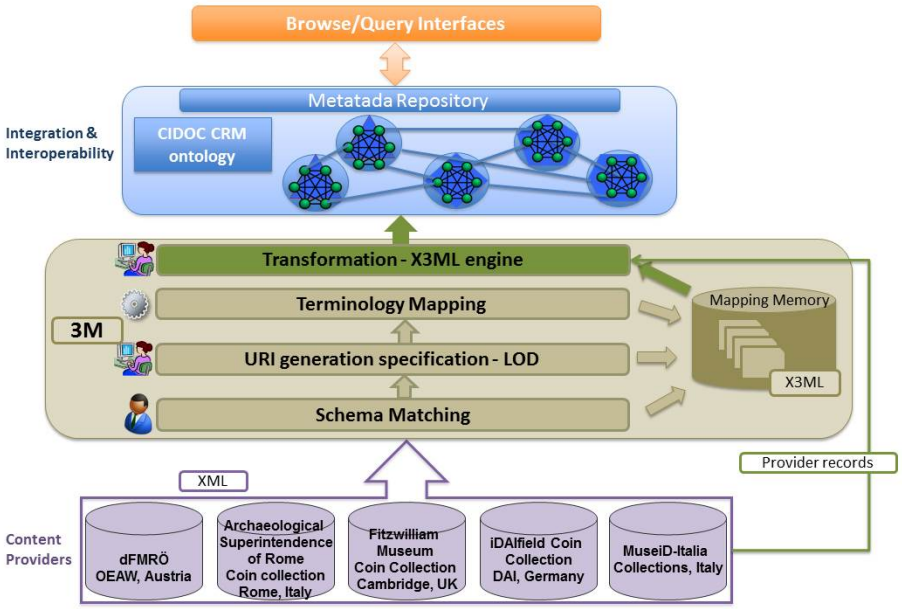


Fig. 1. Aggregation workflow

2. **URI generation specification**: this step aims at defining the functions assigning an appropriate URI to each resource found in the source datasets. Domain experts contribute to this step their expertise on namespaces as well as any policy on naming that is in place in the institution where the integrated dataset will be deployed; the task of IT experts is to properly configure the tools so that the chosen URIs are generated.
3. **Terminology mapping**: this step produces mappings from the thesaurus used by each source dataset to a common thesaurus that is used by the agreggator database. It is similar to the schema matching step, and requires the same tight collaboration amongst different experts.
4. **Transformation**: this is the final step that transforms every record of the source dataset to a set of appropriate RDF triples, subsequently stored in the resulting integrated dataset.

The datasets that we adopt in the ARIADNE integration use case have several differences concerning the origin, language used, purpose of creation and use. Having been created by various institutions and for different purposes, they have quite different data structure, despite the similarity in content. The databases of the SAR and the dFMRÖ, for instance, were created with the purpose of documenting archaeological discoveries which occurred during excavations or surveys and contain many fields reporting information on provenance and discovery conditions. FWM, on the other side, is a museum database whose sole purpose is

to catalog acquisition and inventory data of objects owned and stored by the museum itself, regardless of the archaeological provenance conditions.

The mapping of the coin datasets started with the dFMRÖ archive which was chosen as the first hands-on exercise to map a relational data base schema to CIDOC CRM, since it represents a large class of well-defined traditional databases.

In close cooperation with the domain experts we tried to identify information that was implicit, hidden in forms, hidden in user interface fields or was known only by them. A detailed description of the mapping of the dFMRÖ archive to CIDOC CRM was presented in the CAA2015 conference [DTAM15]. The dFMRÖ mapping was used as a guide for the mapping of the SAR and FWM datasets. The records of the FWM archive contain fields with condensed information that needs to be preprocessed and normalized before it can be mapped to CIDOC CRM. For example, all the information concerning the dimensions of a coin (height, width, weight) is encoded in one field:

```
<Dimension>image(height), 22, mmimage(width), 20, mmweight, 3.74
</Dimension>
```

and needs to be normalized before the actual transformation takes place.

It is worth mentioning that the mapping of a schema to CIDOC CRM is not necessarily unique. There may be different ways of approaching the problem, all correct. However, what is individually correct may turn out to be problematic if considered in the context of a larger process. For example, in the dFMRÖ and SAR mappings, the coin denomination was mapped to a **E55 Type** while in the iDAIfield it was mapped to a **E54 Dimension**. Conceptually both approaches are correct, but their coexistence in the same process is clearly problematic. Rather than imposing a unique style, we have chosen to reconcile such differences at the query level.

The dates are also a crucial point in the integration of the datasets. Different formats and approaches may have been used to encode temporal information in the source databases. To mention just a simple issue, the value *zero* is used as a date in some of the datasets, possibly with different meanings. For instance, such a value might indicate the year in which Jesus was born, or the fact that the year is unknown, or not recorded. This poses several problems. First, *zero* is not a valid date in RDF (or in the underlying XML type system), so the value has to be transformed into a valid date. But in order to carry out this transformation, it is important to clarify the semantics of the zero value in each dataset.

4 Description of the Integrated Infrastructure Set Up for Item Based Management

The ultimate goal of the integration of the diverse coin datasets is to create an environment where users will be able to specify queries that will be evaluated on the common aggregated repository and will be able to combine results coming from the different datasets. The ARIADNE portal will provide a main access

point to integrated repository and an intuitive user interface will guide the user to formulate his query, browse the results and refine the search with facet view. We plan to implement a query interface that will take advantage of the principles of the Fundamental Categories and Fundamental Relatioships as defined in [TD12], [TDTF13].

Currently our work is focused on determining the type of research questions that we would like to support. We have identified the following:

- **Origin** - Where does this coin come from?
- **Tracking** - How did it arrive here?
- **Chronology** - First/last appearance
- **Practical/symbolic value, incidents** - Why is it deposited here?
- **Political message** - Why was it produced (i.e. "minted")?
- **Economic stability, power** - Why was it widely used / not used?
- **Statistics** - Material versus nominal value

There exist several queries that are trivial to be answered by each dataset separately, however they become important if they can be answered by the aggregated repository:

- Find coins minted in the same place/area or by the same authority
- Find coins produced in the same period or time span (typically the same century or half/quarter century)
- Find coins having common shape/iconography/inscriptions
- Find coins made by a specific material.

Combinations of the above queries can be found useful by the researchers of the numismatic community and our first experiments with such queries on the aggregated repository are quite promising. Our experimental aggregated repository contains 72 records (all Roman coins) of the dFMRO archive, 627 records (all Roman coins) of the SAR archive, 517 records (12 Roman coins 1 empty record) of the Pergamon archive (iDAIfield) and 1 record from MuseiD-Italia. The results of some simple queries can be seen in the following table:

Query	Total	dFMRO	SAR	Pergamon	MuseiD
Find all coins	1216	72	627	516	1
Find Roman coins	711	72	627	12	
Find bronze assarius	82	29	52	1	
Find bronze coins	676	50	270	355	1
Find bronze sextans	47		46		1
Find coins produced in -32	22	4	18		

5 Conclusions and Further Work

The activities carried out so far have shown that datsets of different origin, language, property, and of heterogeneous information content, can be successfully

integrated by relying on an ontology that is adequate to capture in conceptual terms the real nature and meaning of the objects described in these datasets. Although the relative homogeneity of the coin class of objects has made much easier the mapping and conversion work, the validity of the methodological approach is universal for any type of archaeological object. CIDOC CRM has proven to be a particularly able ontology to express the conceptual meaning of archaeological entities. However some issues remain still open, such as, for example, the design and implementation of appropriate and efficient user interfaces able to view and query semantically integrated archives like the one we implemented in this case study. The ARIADNE Portal, still under development, is already in the process of providing satisfactory answers to such questions. Once released it should provide all the necessary functionality for querying information thorough all the archaeological archives, regardless of their level of integration. Beside the wide use of CIDOC CRM, there are important domain specific ontologies, where further work will concentrate on a showcase mapping between the numismatic specific ontology Nomisma.org and the generic, global ontology CIDOC CRM. As a result it will be possible to integrate further datasets, which are already using the numismatics ontology Nomisma.org. Future activities would build on the results achieved so far to try to extend the methodology used for the archives of coins to other archaeological archives part of the ARIADNE project, to define the depth at which the integration could be achieved. The release of the CRMarchaeo archaeological extension, expected by the end of the project, will surely simplify the implementation of an interoperability framework at item level.

6 Acknowledgements

We would like to thank K. Vondrovec, E. Aspöck and A. Masur of the Austrian Academy of Sciences for providing access to dFMRÖ and for their help with the mapping. We would also like to thank Sara di Giorgio of MIBACT-ICCU for providing access to the MuseiD-Italia collections.

This work has received funding from the FP7-INFRASTRUCTURES-2012-1 EU project ARIADNE (Grant agreement no: 313193).

References

[BV78] H. Bödefeld and O.v. Vacano. Elektronische Datenverarbeitung in der antiken Numismatik. Ein Projekt zur Erfassung griechischer Münztypen am Althistorischen Institut der Universitt Düsseldorf. *Chiron*, 8:587–604, 1978.

[CRM15] Definition of the cidoc conceptual reference model version 6.1. Available from: http://www.cidoc-crm.org/docs/cidoc_crm_version_6.1.pdf, June 2015.

[DFdJ+15] M. Doerr, A. Felicetti, G. de Jong, K. Konsolaki, B. Norton, D. Oldman, and Wikman T. Theodoridou, M. The SYNERGY reference model of data provision and aggregation. Available from: http://www.cidoc-crm.org/docs/SRM_v1.4.pdf, June 2015.

[dFM07] dFMRÖ - digitale Fundmünzen der Römischen Zeit in Österreich. Available from: http://www.oeaw.ac.at/antike/index.php?id=358, 2007.

[DTAM15] M. Doerr, M. Theodoridou, E. Aspöck, and A. Masur. Mapping archaeological databases to cidoc-crm. In *Proc. 43rd Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2015 SIENA)*, Siena, Italy, April 2015. Archaeopress.

[MMK+15] N. Minadakis, Y. Marketakis, H. Kondylakis, G. Flouris, M. Theodoridou, G. de Jong, and M. Doerr. X3ML framework: An effective suite for supporting data mappings. In *Proceedings of the EMF-CRM2015 workshop (TPDL2015, Poznan)*, August 2015.

[ODdJ+14] D. Oldman, M. Doerr, G. de Jong, B. Norton, and T. Wikman. Realizing lessons of the last 20 years: A manifesto for data provisioning & aggregation services for the digital humanities. *D-Lib Magazine*, 20(7/8), July/August 2014. doi:10.1045/july2014-oldman.

[TD12] K. Tzompanaki and M. Doerr. Fundamental categories and relationships for intuitive querying CIDOC-CRM based repositories. Technical Report TR-429, ICS-FORTH, April 2012.

[TDTF13] K. Tzompanaki, M. Doerr, M. Theodoridou, and I. Fundulaki. Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories. Online Proceedings for Scientific Workshops, 2013.

Appendix: Samples of records from the four datasets

iDAI.field sample record

```
<Aufbewahrungsort>Grabungshaus, Depot</Aufbewahrungsort>
<Auto_Objektkennung>PE08 So 02 - KFN 0002</Auto_Objektkennung>
<Beschreibung>BMC (Mysien) S. 128-129, Nr. 150-157</Beschreibung>
<Erhaltung_Durchmesser>174</Erhaltung_Durchmesser>
<Erhaltung_Gewicht>4,03</Erhaltung_Gewicht>
<Funddatum>09.08.2008</Funddatum>
<Herkunft>Archol. Befund</Herkunft>
<Kampagne>2008</Kampagne>
<KurzbeschreibungMuenze>hellenistische Mnze</KurzbeschreibungMuenze>
<Lage_Details>Auffllung, durch byz. Grber  gestrt</Lage_Details>
<Metall>Bronze</Metall>
<Muenzstaette>Pergamon</Muenzstaette>
<Nominal>Einer (Chalkus)</Nominal>
<Nummer_Fund>2</Nummer_Fund>
<Praegeherr>stdtische Prgung</Praegeherr>
<PS_MuenzeID>49005</PS_MuenzeID>
<Rckseite_Beischrift>[A]KHIO[Y  THPO]</Rckseite_Beischrift>
<Rckseite_Freitext>Schlangenstab</Rckseite_Freitext>
<Rckseite_Motiv></Rckseite_Motiv>
<Stempelstellung>6</Stempelstellung>
<Vorderseite_Beischrift>keine</Vorderseite_Beischrift>
<Vorderseite_Freitext>Belorbeer. Asklepioskopf<Vorderseite_Freitext>
```

The FWM archive sample record

```
<Coin>
 <Accession>Object Number: CM.YG.1008-R(Coins and Medals)</Accession>
 <Acquisition>bequeathed; 1936-07-07; Young, Arthur W.</Acquisition>
 <AlternativeNumber>RRC; 452/2ordering; RR-2672</AlternativeNumber>
 <Category>coin</Category>
 <Collection>Young Collection</Collection>
 <Date>48-07-13 B.C.  47 B.C.</Date>
 <Dimension>
 image(height), 22, mmimage(width), 20, mmweight, 3.74
 </Dimension>
 <Inscription>
 design; obverse; female head wearing wreath and diadem; behind,
 IITdesign; reverse; Trophy with Gallic shield and carnyx; on r.,
 axe; below, CAE SAR
 </Inscription>
 <Maker>
 Caesar, Gaius Julius; moneyer; Roman, 100-44 B.C.Rome; place of use
 </Maker>
 <Material>silver</Material>
 <Name>Roman Republic; Seriesdenarius; denomination</Name>
 <PermanentIdentifier>
 http://data.fitzmuseum.cam.ac.uk/id/object/114778
 </PermanentIdentifier>
 <ProductionNote>mint moving with Caesar</ProductionNote>
 <ProductionPlace>Rome, place of use, state</ProductionPlace>
 <Technique>struck</Technique>
 <obverseImg>
 http://www-img.fitzmuseum.cam.ac.uk/img/cm/cm7/CM.YG.1008-R(1).jpg
 </obverseImg>
 <reverseImg>
 http://www-img.fitzmuseum.cam.ac.uk/img/cm/cm7/CM.YG.1008-R(2).jpg
 </reverseImg>
</Coin>
```

The SAR archive sample record

```
<coins_sar>
  <ID>2680</ID>
  <inv>1812-1</inv>
  <weight>38.36</weight>
  <diam_min>36.5</diam_min>
  <diam_max>37</diam_max>
  <posit>0</posit>
  <authority>Anonimo</authority>
```

```
  <metal>AE</metal>
  <mint>Roma</mint>
  <nominal>As</nominal>
  <class>repubblicane</class>
  <car_fisiche>integra</car_fisiche>
  <regio_naz>Latium</regio_naz>
  <crono>secolo</crono>
  <from_year>-225</from_year>
  <to_year>-201</to_year>
  <century>III a.C.</century>
  <part_century>ultimo qua</part_century>
  <calc_century>III a.C.</calc_century>
  <armadio>1</armadio>
  <cassetto>3</cassetto>
  <d_leggenda>ANEPIGRAFE</d_leggenda>
  <r_leggenda>NON REGISTRATA</r_leggenda>
</coins_sar>
```

The dFMRÖ archive sample record

```
<COIN>
<ID>626</ID>
<COUNTRY_ID>1</COUNTRY_ID>
<FIND_SPOT_ID>242</FIND_SPOT_ID>
<FIND_MANNER_ID>2</FIND_MANNER_ID>
<FIND_DATE>-</FIND_DATE>
<AUTHORITY_ID>565</AUTHORITY_ID>
<ISSUER_ID>243</ISSUER_ID>
<DENOMINATION>239</DENOMINATION>
<MINT_ID>2291</MINT_ID>
<OFFICINA>99</OFFICINA>
<DATE_CA>1</DATE_CA>
<DATE_FROM>-100</DATE_FROM>
<DATE_TO>0</DATE_TO>
<DAT_VAL>1090010001</DAT_VAL>
<WEIGHT>0.43</WEIGHT>
<DIE_AXE>9</DIE_AXE>
<STATUS_ID>1</STATUS_ID>
<ARCH_INFO>neben der Bundesheerkaserne</ARCH_INFO>
<PH_NAME>000626</PH_NAME>
<DAT_TXT>ca. 100 v. - 0 n. Chr.</DAT_TXT>
</COIN>

<AUTHORITY>
<AUTH_ID>565</AUTH_ID>
<AUTH_NAME>KELTEN</AUTH_NAME>
```

```
<AUTH_NAME_EN>CELTS</AUTH_NAME_EN>
<AUTH_ORDER_KEY>20</AUTH_ORDER_KEY>
<AUTH_S_ORDER_KEY>000020</AUTH_S_ORDER_KEY>
</AUTHORITY>

<COUNTRY>
<COUNTRY_ID>1</COUNTRY_ID>
<COUNTRY_NAME>sterreich</COUNTRY_NAME>
<COUNTRY_NAME_EN>Austria</COUNTRY_NAME_EN>
</COUNTRY>

<DENOMINATION>
<DEN_ID>239</DEN_ID>
<DEN_NAME>Kls</DEN_NAME>
<DEN_ORDER_KEY>110</DEN_ORDER_KEY>
<DEN_S_ORDER_KEY>000010_000020_000110</DEN_S_ORDER_KEY>
<DEN_METAL>2</DEN_METAL>
<EXCLUDE>0</EXCLUDE>
</DENOMINATION>

<FIND_MANNER>
<FM_ID>2</FM_ID>
<FM_NAME>Streufund</FM_NAME>
<FM_NAME_EN>stray-find</FM_NAME_EN>
</FIND_MANNER>

<FIND_SPOT>
<FS_ID>242</FS_ID>
<FS_NAME>Mautern</FS_NAME>
<FS_ORDER_KEY>10</FS_ORDER_KEY>
<FS_S_ORDER_KEY>000020_000030_000090_000170_000010</FS_S_ORDER_KEY>
<EXCLUDE>0</EXCLUDE>
</FIND_SPOT>

<ISSUER>
<PR_ID>243</PR_ID>
<PR_NAME>Boier</PR_NAME>
<PR_NAME_EN>Boii</PR_NAME_EN>
<PR_ORDER_KEY>20</PR_ORDER_KEY>
<PR_S_ORDER_KEY>000020_000020</PR_S_ORDER_KEY>
<EXCLUDE>0</EXCLUDE>
</ISSUER>

<MINT>
<MINT_ID>2291</MINT_ID>
```

```
<MINT_NAME>kelt. Mzst.</MINT_NAME>
<MINT_NAME_EN>celtic mint</MINT_NAME_EN>
<MINT_ORDER_KEY>100</MINT_ORDER_KEY>
<MINT_S_ORDER_KEY>000020_000100</MINT_S_ORDER_KEY>
<EXCLUDE>0</EXCLUDE>
</MINT>

<STATUS>
<STATUS_ID>1</STATUS_ID>
<STATUS_NAME>Mnze</STATUS_NAME>
<STATUS_NAME_EN>genuine</STATUS_NAME_EN>
</STATUS>

<METAL>
<MET_ID>2</MET_ID>
<MET_NAME>AR</MET_NAME>
</METAL>
</dataroot>
```

Integrating Terminological Tools and Semantic Archaeological Information: the ICCD RA Schema and Thesaurus

Achille Felicetti[1], Ilenia Galluccio[1], Cinzia Luddi[1], Maria Letizia Mancinelli[2], Tiziana Scarselli[3], and Antonio Davide Madonna[3]

[1]PIN, VAST-LAB, Prato, Italy
[2]MiBACT-ICCD, Istituto Centrale per il Catalogo e la Documentazione, Rome, Italy
[3]MiBACT-ICCU, Istituto Centrale per il Catalogo Unico, Rome, Italy
{achille.felicetti,ilenia.galluccio,cinzia.luddi}@pin.unifi.it
{marialetizia.mancinelli,tiziana.scarselli,
antoniodavide.madonna}@beniculturali.it

Abstract. This paper describes the process of mapping, translation and publication in SKOS format of the RA Thesaurus, a terminological tool developed by the Italian Ministry of Cultural Heritage (MiBACT) as a part of the official documentation used for the recording of archaeological finds. In particular, the RA Thesaurus is intended to provide unified and meaningful terminology for the description of archaeological objects according to the MiBACT official cataloguing standards. After describing the thesaurus, the logic with which it was developed and its internal structure, we report the various phases of the conversion, both from a theoretical and implementation point of view, and the various technologies used for the publication of the thesaurus on the web. This work is a collaborative effort between PIN and MiBACT carried out under the ARIADNE project.

Keywords: Archaeology, Mapping, Thesauri, ICCD, CIDOC CRM, SKOS

1 Introduction

ARIADNE is a European project focusing on integration of existing archaeological research data infrastructures to enable the use of distributed datasets and services by means of new and powerful technologies as an integral component of the archaeological research methodology. Among other activities, ARIADNE is also actively working on building a coordinated system of multilingual terminology tools able to meet the many needs of the international community of archaeologists. As part of these integration activities, the valuable work of mapping national catalogue schemas on international standards is a critical step; at the same time integration of terminology resources is necessary to overcome linguistic barriers that frequently slow down the integration processes. We have extensively described the process of CIDOC CRM encoding of the RA Schema,

released by ICCD for documenting archaeological artefacts in Italian archaeology, in a previous work [1]. The mapping was presented as work in progress at that time. Since then, new extensions of the CIDOC CRM (and in particular CRM*archaeo*) have been released which are now able to provide more possibilities for the enrichment of the semantic archaeological information and a more archaeological oriented means of documentation. The release of new versions and the creation of new extensions of the CIDOC CRM gave us the opportunity to investigate how the mapping could be improved. This allowed us to bring the mapping to a stage very close to completion, although much work still remains to be done. The RA Schema is closely linked to the RA Thesaurus, a sophisticated vocabulary providing all the necessary terminological facilities for an efficient and well-structured recording of the objects coming from archaeological excavations. The vocabulary has been implemented by ICCD to support the encoding of two specific fields (OGTD - CLS). These two fields describe the definition of the object and its class and production. This paper will focus and propose integration between the RA Schema and its thesaurus, based on W3C recommendations and using numerous tools developed and used by several partners in the ARIADNE project.

2 ICCD and the Standards for Cultural Heritage

ICCD is the Italian Central Institute for Catalogue and Documentation, one of the seven Central Institutes of the Italian Ministry of Cultural Heritage whose main goal is to create a centralized national catalogue of Italian cultural heritage. The activity of the Institute is based on the research and development of tools, methods and standards for knowledge, protection and enhancement of the cultural and artistic heritage in Italy. It mainly provides the management of the national general catalogue of archaeological, architectural, historical, artistic and ethno-anthropological heritage, the development of cataloguing methodologies and standards, and the coordination of the technical institutions involved in the cataloguing activities on the national territory. ICCD also provides tools and best practices for implementing these standards with the clear intent of unifying and streamlining processes related with the cataloguing activities, to guarantee quality and to implement standardisation and interoperability at a national level.

To ensure that this happens efficiently, the Institute creates and releases a series of organic resources and recommendations to support the standardization process in all its aspects. These include detailed regulations describing the various tools and the way they should be used, a set of schemas and forms to collect information in a structured way according to the different asset types, authority files to guarantee homogeneity for the common transversal key concepts and entities, thesauri and terminological tools to provide uniform layers of information and a common language. Among the latter category, one of the most important tools released by ICCD is the RA Thesaurus, a tool providing standard names for the definition of archaeological artefacts described using the RA schema,

the ICCD standard schema used for the recording of movable objects. The RA Schema is the most used and well established standard for Italian archaeology so far. For this reason, ICCD has invested a lot of effort in the definition of a terminological tool able to provide standardized and unambiguous names for specific fields of the schema. The creation of the RA thesaurus is one of the best results of this effort.

3 The ICCD RA - CIDOC CRM Mapping

In the previous work carried out together with ICCD, a detailed analysis of the RA Schema was made to map the most significant model of ICCD archaeological cataloguing system to CIDOC CRM. The RA Schema is used to record movable objects. It is one of the most used for Italian archaeology because of the huge and ever increasing amount of artefacts found during excavations. The RA Schema contains a large number of descriptive information and "cross-sections" allowing cross references with other ICCD resources. The RA Schema, together with the RA Thesaurus, features one of the best tools of this kind in the international panorama of cataloguing systems. The previous mapping work was carried out on CIDOC CRM and took advantage of version 5 of the model, released in 2013. However, in the last two years, a new version and numerous extensions of CIDOC CRM have been released. Version 6 and the CRM*archaeo* [2] and CRM*sci* [3] extensions, much more suitable for the description of archaeological phenomena, have strongly enhanced the representation and mapping of excavation entities. CRM*archaeo*, in particular, is being developed by the ARIADNE project to facilitate the encoding of archaeological entities. Given this, we decided to update the previous mapping in order to provide a stronger archaeology-oriented logic to the various concepts and relationships that the RA Schema presents.

One of the most difficult problems to solve during the previous mapping was the representation of the "finding" event, intended as the excavation activity during which objects are found. This event is of paramount importance in archaeology because it is fundamental to trace the object's provenance and to reconstruct its history. Following the CIDOC-CRM model, we represented the archaeological objects by using the *E22 Man-Made Object* class. However, to describe their relationships with the two important activities of "survey" (corresponding to the "RE" field of the RA Schema) and "excavation" (specified in the "DSC" field), CIDOC CRM core only provided a "change of ownership" relationship that hardly fits here but we decided to use it anyway. Our previous mapping appeared as shown in Figure 1.

Thanks to the release of the new extensions and a deep analysis of the cross-section relating the RA Schema [4], the new mapping now shows a more accurate rendering of these concepts. To express the "object found during an excavation" relationship, CRM*archaeo* provides the *O19i was object found by* property, through which it is possible to link the artefact with the new *S19 Encounter Event* class, expressly designed to render the concept of "finding" as an event which occurred (*P7 took place at*) at a given Site (*E7*) identified by

Fig. 1. ICCD-RA/CIDOC-CRM mapping

a given appellation (*P57 is identified by - E44 Place Appellation*), as shown in Figure 2. This constitutes a more accurate representation of these concepts.

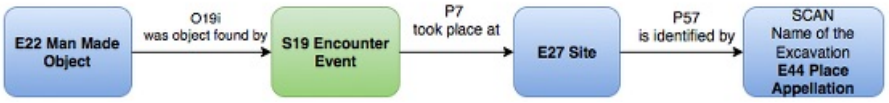


Fig. 2. ICCD-RA/CIDOC CRM/CRM*archaeo* mapping

3.1 The ICCD RA Thesaurus

The RA Thesaurus was developed expressly to provide standardised values for some of the "OG-OGGETTO" (*Object*) fields of the RA Schema. The content of the thesaurus is organized in a tabular structure with five columns arranged according to the hierarchical levels provided by the thesaurus. The first three columns, used to fill the CLS field of the RA schema, present the categories' three levels of hierarchy, to which any concept can belong; column four lists the main terms for the definition of the objects; column five provides specifications of the main terms in accordance with morphological, functional or partitive criteria. Both columns four and five are meant to provide standard terms for the OGTD field of the RA Schema. Additional columns, reporting further attributes and specifications for each term and subterm, such as descriptive notes and sample images, are also present (see Fig. 3). Images are an added value of this tool for their ability to visually show what words are not always able to say. We have already investigated some of the possibilities to encode figures in our mapping, but unfortunately the tools at our disposal do not always allow a clear definition of these entities. For sure it will be important, in future versions of the thesaurus, to define a standard mechanism for associating concepts with their images even in the SKOS version of the thesaurus.

The RA Thesaurus differs from the other terminological tools created by ICCD in the very sophisticated structuring criteria it follows, made more complicated by the large amount of information deriving from Italian archaeology and the huge number of classifications and nomenclatures it provides. In particular, the thesaurus is structured according to a multilevel schema based on concept coordination, a typical KOS activity in which concepts are combined with each other in order to produce meaningful "sentences" that define complex

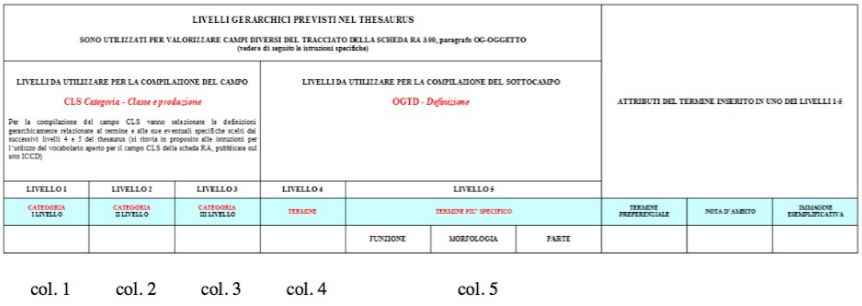


Fig. 3. The ICCD RA Thesaurus model

concepts. Generally speaking, there can be two types of concept coordination: pre-coordination and post-coordination. The key distinction between the two relies on when the actual coordination occurs in relation to an information retrieval event. Pre-coordination is decided and implemented before the information retrieval time, by a KOS maintainer or by an indexer who is using the KOS itself. This occurs, for instance, when an indexer takes two existing concepts from a concept scheme, such as "Coins" and "Mintage", and explicitly combines them with a given syntax, such as "Coins-Mintage", to index a particular document. Post-coordination, on the other hand, is performed as part of an information retrieval task, for instance through a SPARQL query able to retrieve all documents indexed using both "Coins" and "Mintage" concepts [5]. The RA Thesaurus follows the post-coordination approach to create ad hoc concepts by using the elements of a given schema. Each concept is in fact provided with all the necessary subterms depending on it, which can belong to three specific semantic areas according to the specification provided: either functional (i.e. relative to the specific function of the object), partitive (i.e. relative to a specified part of the object) or morphological (i.e. linked to the different forms that from time to time an object may present). The structure of the thesaurus is obviously functional to the specific cataloguing activities. Each concept is thus created on the fly by combining the main terms with all the related subterms required to render the specific name that a concept should show in a given context. Figure 4 provides an example of how the thesaurus is structured by reporting the various facets of the term "cintura" (belt) and its related functional and morphological subterms:

It is evident from the example above that the thesaurus itself does not offer a closed and exhaustive list of all possible terms that can be used during the compilation of the schema. Instead, it is a reference tool that, after a general term is fixed, assists the user in proceeding to further specifications by the addition of suitable subterms to gradually approximate the precise semantic meaning of the object to be described.

The flexibility of this structure allows it to achieve a significant depth of semantics, where required, and to build specific definitions of several types of objects,

Reference Term	Specific Terms	Scope notes
cintura		
	a fascia	MORPHOLOGY
	a losanga	MORPHOLOGY
	a placche	MORPHOLOGY
	affibbiaglio	PART
	borchia	PART
	multipla	MORPHOLOGY
	per la sospensione delle armi	FUNCTION
	puntale secondario	PART

Fig. 4. Example of the thesaurus structure

including those in fragmentary conditions (for instance by means of the partitive subconcepts).
This will overcome the necessity to define in advance the entire terminological apparatus suitable to describe the infinite variety of situations the archaeologists may face.
Just to remain with the example above, from a logical point of view, if an archaeologist finds a stud (*borchia*) pertaining to an ancient belt (*cintura*) intended for the suspension of a sword or other similar weapons (*per la sospensione delle armi*).
A valid definition would be composed as follows:
Cintura (main term) +
per la sospensione delle armi (morphological aspect of the main term) +
borchia (part of the object that was found)

in order to have an entry like this:

Cintura per la sospensione delle armi, borchia (Belt for weapons suspension, stud)

representing an exhaustive explanation of the fragmentary object itself and of the bigger object which is part of, and also as a valid entry from the terminolog-

ical point of view following the formal recommendations provided by the ICCD guidelines and validation systems.

3.2 A SKOS Mapping Proposal

SKOS is the standard chosen by the ARIADNE project for the encoding of all terminological resources to be used in its integration plan, and for the undeniable advantages provided to integration and interoperability by its RDF-based format. As one can easily understand from what was previously stated, the "combinatorial" nature of the RA Thesaurus, and especially of the sections intended for the encoding of the OGTD field (column four and five), makes it very difficult to encode in a SKOS compatible format, which requires that a complete, self-consistent and self-sufficient definition exists in the thesaurus for each item or concept. The SKOS vocabulary itself does not provide any mechanism for expressing that a given concept may consist of other pre-coordinated concepts. It is, of course, possible to extend SKOS to establish a pattern for representing coordinated concepts, for instance by stating a new sub property, as in the following example:

iccd:coordinationOf a rdf:Property;
rdfs:domain skos:Concept;
rdfs:range rdf:List.

and then use the new property this way:

iccd:coinsMintage a skos:Concept;
iccd:coordinationOf (iccd:coins iccd:mintage);
skos:prefLabel "Coins-Mintage"@en.

However, patterns for pre-coordination have not yet been exploited by the SKOS community and solutions of this kind have not been explored fully enough to warrant their inclusion in the official SKOS vocabulary. Analyzing the RA Thesaurus, PIN and ICCU identified a possible solution. We tried to follow a different approach, more "pre-coordination oriented" to rearrange, where possible, the original content according to semantic criteria in order to define meaningful self-consistent concepts in the SKOS representation. After discussing the matter in depth, we proposed the following solutions:

1. The partitive specification subterms are in many cases independent terms related with the main term mostly by a part-whole relationship. Thus, it is possible to describe this relationship by using the *skos:narrowerPartitive* property to define them. This is particularly suitable if we consider that the same partitive term could occur for different main concepts: both a belt and a flag could have a *puntale* (ferrule) as partitive concept. Therefore, it is important to clearly define the hierarchy of these kinds of objects. Alternatively, it would be possible to combine main terms with their partitive terms in order to define complete and

self-consistent concepts, to be then defined as narrower terms of the main ones. In the previous example, we could define, for instance, a new *puntale di cintura* (belt ferrule) term, which would be clearly distinguished by a *puntale di insegna* (flag ferrule), the two being totally different, although very similar, objects.

2. The morphology and functional specification subterms are meaningless in themselves. They become meaningful only when combined with their main term. Creating SKOS narrower terms from these elements requires, for each morphological or functional term, the creation of a subterm obtained by combination with the super concept, in order to obtain a set of semantically consistent narrow terms. There is no semantic meaning in *multipla* itself unless this concept is used together with *cintura* in order to specify, in this case, the typology of a given belt. *Cintura multipla* is, on the other hand, a perfectly consistent concept.

Multiple combination of partitive, morphological and functional sub concepts to create specific entries, even if not impossible, would be very difficult to implement in SKOS due to the exponential growth of all possible combinations. At present, we propose not to extend the pre-coordination operations beyond the minimum requirements of semantic understandability and to use more than one SKOS concept to describe specific archaeological objects if required.

4 SKOS Encoding of RA Thesaurus

From a technological point of view, the RA Thesaurus was created starting from 2008 on the basis of the terms extracted from the database maintained by the "Sistema Informativo Generale del Catalogo" (SIGEC). Its development also went through various phases of data cleaning and strengthening. The RA Thesaurus is currently an "open vocabulary", meaning that it is not meant to have a stable form since its content can be updated and modified by ICCD during further stages of research. Currently, the available version of the vocabulary is in textual format that is organized in a tabular structure, whose fields comply with the ISO standard norms for thesauri. In order to make the original textual information interoperable and ensure integration with semantic terminological tools, it was necessary to encode them in a structured and standard format.

The process we implemented for the SKOS encoding of the RA Thesaurus is a proposal for its *re-engineering* as a formal ontology and for making the knowledge it provides explicit in a formal sense. The whole process of encoding required a set of subsequent steps for data analysis, adjustment, conversion, publication and enrichment, in which the original textual data has been processed using both open source tools and *ad hoc* scripts.

The process can be subdivided into two analytic phases (see Fig. 5):
1. In the first analytic phase we focused on encoding the key fields of the original thesaurus, such as concepts and classes. The result of the first phase consisted in the creation of a SKOS/RDF version of the RA Thesaurus obtained through the mapping between the main concepts and the SKOS Core Vocabulary.
2. In the second phase, we focused on the integration of all morphological, functional and partitive aspects related to thesaurus' concepts. The analysis of this

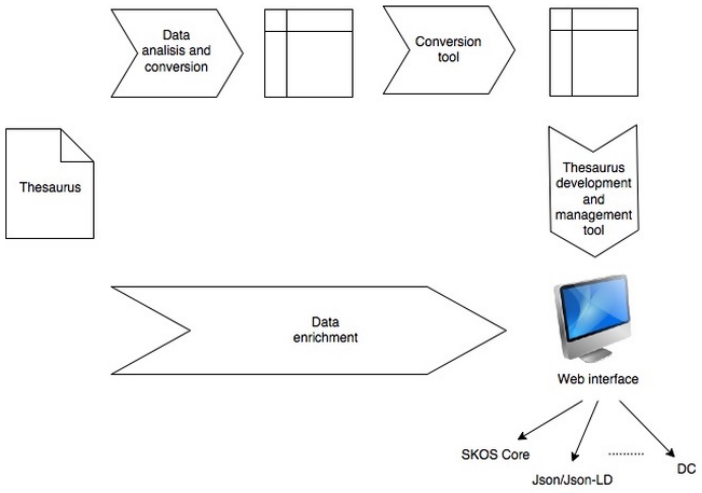


Fig. 5. SKOS encoding process

additional information required further investigation into how SKOS extensions could be used for the publication of thesauri in a semantic format.

4.1 Thesaurus Conversion Using SKOS Core Vocabulary

The conversion of the RA Thesaurus initially required a deep data analysis to define a precise mapping between its main fields and the SKOS Core Vocabulary [6] in order to use its set of properties and classes to express the conceptual content of the thesaurus as an RDF graph. The fields examined in the first analytic phase are levels one and two, containing categories and subcategories, and level four, containing the main terms for the description of the artefacts. With reference to level five, we limited our analysis to the functional facet only and we considered the descriptive notes in the attribute fields. Classes and terms were mapped using the *skos:Concept* entity, main terms were mapped as *skos:prefLabel*, non-preferential terms as *skos:altLabel*, notes were encoded using *skos:scopeNote*. The *skos:broader* and *skos:narrower* properties were used to express the hierarchical relationships between categories or concepts. The functional specification of a term was expressed through the *skos:narrower* relation with a subterm obtained by combination with the super concept.

Figure 6 shows an example of the mapping expressed by using SKOS entities. Each concept coming from the RA Thesaurus is represented by a blue circle. The central circle depicts the concept of Cintura (belt) while the red circle represents the thesaurus itself. Arrows connecting the various circles represent the SKOS relationships existing among them. The mapping definition on the SKOS Core Vocabulary was followed by the use of an *ad hoc* script and of a specific tool that allowed the conversion of a huge textual file into RDF format.

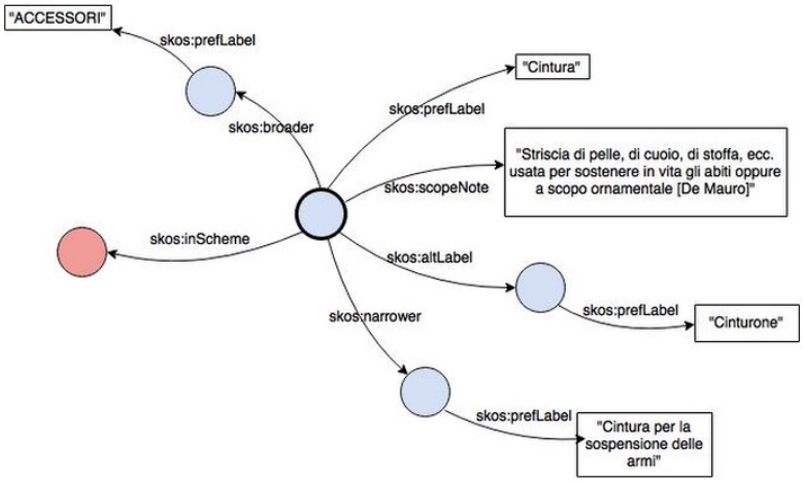


Fig. 6. Example of mapping expressed by using SKOS entities

At first, the original thesaurus was manipulated and converted in order to create a CSV file that satisfied some specific technical requirements. The script was developed in Perl language and was intended to select a specific thesaurus' subset of fields, to sort and to clean the information and to convert them into a custom CSV file. Subsequently, the Stellar Console tool was applied to further elaborate this file. Stellar Console is an open source command line utility application developed in the framework of the AHRC-funded project "Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources" (STELLAR) [7]. The Console accepts input format such as CSV in order to produce a more structured output such as SKOS/RDF or CIDOC-CRM/RDF by applying a set of customizable templates. The templates look for the presence of particular field names in the input data, and process each row in turn using the values contained in these fields. The use of the conversion feature from the custom CSV files to the SKOS/RDF of Stellar Console is the final step for the conversion of the main subset of RA Thesaurus from a textual format to a structured, semantic and interoperable format.

4.2 Thesaurus Publication and Enrichment Using SKOS Extensions

In the second analytic phase, the publication of the thesaurus was analysed and tested on a vocabulary server. Possible solutions for mapping and integrating the fields that were not converted in the first analytic phase were consequently studied and tested.

In order to produce the necessary results for the RA Thesaurus publication, it was important to consider two fundamental aspects. The first was a vocabulary web server supporting international standards such as SKOS and the ISO

thesaurus norms; the second was a vocabulary web application which supports multilingualism, semantic thesauri and data enrichment. All these aspects, in our opinion, are fundamental to make the RA Thesaurus even more flexible for future study phases by expanding and integrating it with further multiple extensions.

We considered different possibilities to achieve the above-mentioned results, by choosing TemaTres as the most pragmatic solution. TemaTres is an open-source, web-based thesaurus management package [8] that supports the handling of vocabularies in accordance with the ISO standard thesaurus norms, including the last ISO-25964 [9]. The main features of TemaTres include a functional user interface for editing and browsing, good search capabilities, and the ability to export all or part of the thesaurus in a number of standardized forms (Json, Json-LD, SKOS Core, DC etc.). TemaTres easily allows data import in SKOS/RDF format and some of the more advanced features include the ability to link terms between two different vocabularies. A test version of TemaTres was installed on a local server and used to import the SKOS/RDF thesaurus version containing the main concepts in order to proceed further with the enrichment work. The TemaTres publication of the RA Thesaurus provides many editing and search facilities. One of the most important is the ability to customize and automatically generate URIs used to unambiguously identify and reach resources from any context. For generating suitable URIs we have used - by means of testing - the official ICCD namespace (http://www.iccd.beniculturali.it), which will be useful for the future installation of TemaTres on the ICCD server and for the creation of consistent and unambiguous URI/URL to make the RA Thesaurus available also in a Linked Open Data format. The conversion of the fields related to morphological and partitive specification of terms required further actions on the data. We used the TemaTres administration facilities for this semantic enrichment. We mapped the morphological specifications using *skos:broader* and *skos:narrower* properties. The partitive specification subterms was mapped using the last ISO standard on thesauri ISO 25964[10]. One of the innovations introduced by the current norm is the possibility to make explicit the nature of semantic relationships, in particular we focused on the changes regarding the hierarchical relationships. To extend the richness of thesauri, the SKOS Core hierarchical relationships depicted through the tags BT and NT can be further divided into generic (BTG/NTG), partitive (BTP/NTP) and instantial (NTI/BTI). ISO 25964 specifies that this relationship holds "between a pair of concepts when the scope of one of them falls completely within the scope of the other" [11].We introduced the BTP and NTP relationships using the corresponding property in the 'iso-thes' namespace: *iso-thes:broaderPartitive* and *iso-thes:narrowerPartitive* [12]. The example in Figure 7 shows that a "fibbia di cintura" (belt buckle) concept stated this way, for instance, specifies that the fibbia is part of a "cintura" (belt), whereas a "fibbia" (buckle) per se could also be part of other objects, for instance, a weapon, a garment and so forth. Therefore, the BTP/NTP relationships cannot be automatically inferred by the subconcept only because it could be part of many objects.

The image field of the RA Thesaurus is also a very interesting case. As already mentioned, images increase the richness and meaningfulness of concepts, their presence being sometimes crucial, especially in cases where proper understanding of the archaeological objects may remain ambiguous. In a 2005 version of the SKOS Core Guide W3C Working Draft [13], the Working Group proposed the use of symbolic labels, as part of the labelling properties, to label a concept with an image. Symbolic labels could be used to assign preferred and alternative symbolic labels to a concept by means of the *skos:prefSymbol* and *skos:altSymbol* properties. This solution would have been the most appropriate for the mapping of the RA Thesaurus sample image, but in the subsequent W3C Recommendation [14], symbolic labelling elements were removed, although no explicit deprecation axioms were expressed in the schema. In order to achieve a publication of the thesaurus that complies as far as possible with the W3C specifications, we preferred not to use the solution proposed in the SKOS Core Guide W3C Working Draft, but to use the current W3C Recommendation only. According to the latter, sample images can be regarded as accessorial information of the SKOS concepts. The relationship can be mapped using the *skos:note* property, considering that there is no restriction on the nature of the information that the property can associate with the concept.

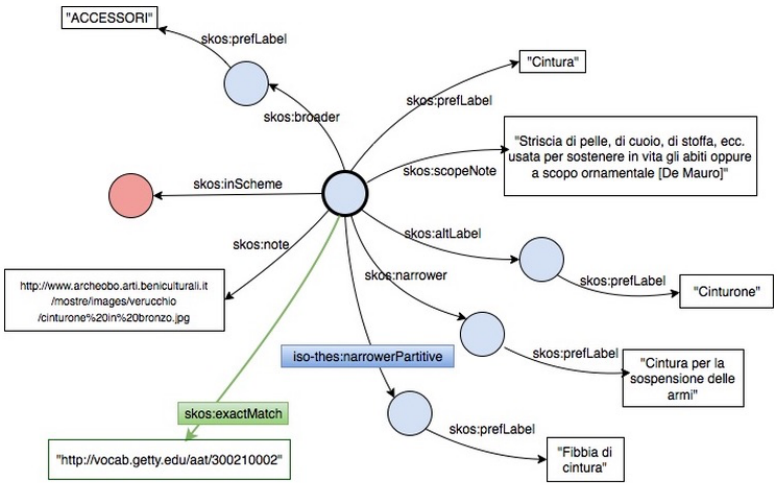


Fig. 7. Example of mapping expressed by using SKOS extensions

5 Getty AAT Mapping

The AAT thesaurus (Art Architecture Thesaurus - Getty Institute) [15] was chosen by ARIADNE to represent a common spine and to constitute a facet

allowing search and faceted browsing across all the terminological tools that the project is collecting. Integration will be based on mappings of national/local vocabularies to the AAT thesaurus. This will allow interoperability over the subject metadata in different partner languages via the common AAT spine. The issue of multilingualism is a matter that needs to be taken into account, not only because of the variety of national thesauri that are going to be integrated by the ARIADNE initiative, but also for the future creation of common and transnational terminological tools. Linguistic issues often make the direct mapping of a concept via the *skos:exactMatch* property on AAT concept difficult, but hopefully the most significant issues will be resolved by the end of the project. The conceptual mapping between the ICCD RA Thesaurus and AAT has been completed and revised; for this purpose it was decided to manually construct a mapping from the various terms and functions (if any), following in sequences the three main categories of the RA Thesaurus. The work pattern was based on an Excel representation of the thesaurus to which additional columns were added in order to specify:

- the *targetLabel* and the unique identifier (ID) of the corresponding definition/term selected in AAT;
- the SKOS schema properties (*skos:closeMatch*; *skos:exactMatch*; *skos:broadMatch* and *skos:matchURI*);
- the name of the institution in charge of the definition of each specific mapping (creator).

Only a subset of the RA Thesaurus was taken into account to demonstrate the feasibility of these operations. The subset includes one thousand, one hundred and ninety one terms related to ten major categories (highlighted in the original source as "livello_1_categoria ") relating to:

- CLOTHING AND ACCESSORIES
- FURNISHING
- TRANSPORTATION
- CONSTRUCTION INDUSTRY
- PAINTING
- ARCHAEOBOTANICAL FINDINGS
- ARCHAEOZOOLOGICAL FINDINGS
- SCULPTURE
- INSTRUMENTS - TOOLS AND OBJECTS OF USE
- GENERAL TERMS

The analysis for finding the corresponding entries in the AAT thesaurus took into account the information provided by scope notes and images accompanying each concept; extensive web searches were performed to find the most appropriate matching term between Italian and English; and terminological researches was carried out using different resources to identify synonyms to make the associated *targetLabel* as unique and as precise as possible.

The mapping work has identified :

- 457 *broadMatch* associations
- 104 *closeMatch* associations
- 630 *exactMatch* associations

Three examples of association are provided in the following table:

Categoria						
livello1	**livello2**	**livello3**	**Livello4 termine**	**targetlabel**	**ID**	**matchlabel**
Mezzi di trasporto	Terrestri	A trazione animale	**cisium**	two-wheeled carriages	300215685	broad match
Strumenti - Utensili e Oggetti d'uso	Armi e Armature	Armi da difesa	**farsetto da armare**	arming doublets	300226824	close match
Scultura			**imago clipeata**	**clipei (portraits)**	300178246	exact match

Fig. 8. Example of RA/AAT associations

At the end of the mapping work we can say that the most significant activity, from the scientific-methodological point of view, has been the review of the whole process. Started as punctual control "1: 1" of correspondence between the terms of the two terminology tools (thesaurus ICCD / RA and AAT), this review has, in fact, been expanding by realizing the mapping of the terminological categories relating to individual entries with the codes referring to the facet and the hierarchy AAT. This has made possible:

1. disambiguating and correcting matches previously selected - and often lexically corrected - but decontextualized from the original domain of belonging;
2. providing the basis for a future matching job between different categories of multilingual thesauri.

It is worth underlining that the focus of the whole work of mapping is the concept of individual terms meant as records entered in a complete hierarchical structure of related terms and notes. Among the results which have been achieved - and which are highlighted though the mapping between classes - we can state the high level of correspondence between the ICCD/RA thesaurus entries and the AAT Thesaurus record types. Out of one thousand, one hundred and ninety one basic records one thousand, one hundred and sixty four among them are linked to "concept" and only twenty seven to "guide term". According to the AAT Thesaurus guidelines:

– Concept: Refers to records in the AAT that represent concepts; records for concepts include terms, a note, and bibliography.
– Guide term: Refers to records that serve as place savers to create a level in the hierarchy under which the AAT can collocate related concepts. Guide terms are not used for indexing or cataloguing.

6 Conclusions and Further work

The study and analysis of the RA Thesaurus allowed us to fully understand the complexity of the challenges arising from the need to define, by means of standard nomenclatures, objects of such various and multifaceted nature as archaeological objects are. The ICCD RA vocabulary, being the result of years of research by a team of experts in the field of Cultural Heritage, is definitely an irreplaceable resource that adequately meets this need. Its structure is certainly an important point of arrival on the road to standardization. From a methodological point of view, the work carried out has highlighted both conceptual and procedural challenges that arise when attempts are made to handle a complex structure in a standard tool. The results achieved so far are considered satisfactory, also in consideration of the fact that the work is at an intermediate stage and that further studies and investigations will be necessary before the conversion of the entire thesaurus can be completed. Future activities will include a clear and unambiguous definition of complex concepts, such as those arising from the combination of multiple terms and subterms; and the definition of precise criteria for the inclusion of images, which, as stated, is one of the distinctive features of this vocabulary. The choice of AAT as the common standard partially solves the multilingualism issues, providing labels in different languages for the terms already mapped. We must instead provide appropriate translations for those that have no equivalent in the thesaurus of the Getty Institute. At the end of the ARIADNE project, the RA Thesaurus will become part of the rich set of terminological tools that the project is already collecting in order to integrate them into the platform on which real interoperability will take place. The ARIADNE Portal will make this resource available and easily accessible online for external use outside of the project. The publication as Linked Open Data, also provided by the project, will guarantee its availability in other Cultural Heritage scenarios.

7 Acknowledgements

The present work has been supported by the ARIADNE project, funded by the European Commission (grant 313193) under the FP7 INFRA-2012-1.1.3 call. The authors opinion do not necessarily reflect those of the European Commission

References

1. Felicetti, A., Scarselli, T., Mancinelli, M.L., Niccolucci, F., (2013) Mapping ICCD Archaeological Data to CIDOC-CRM: the RA Schema, Vladimir Alexiev, Vladimir

Ivanov, Maurice Grinberg (eds.): Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), Valetta, Malta, September 26, 2013, CEUR-WS.org/Vol-1117, pp 11-22

2. CRMarchaeo http://www.ics.forth.gr/isl/index_main.php?l=e&c=711
3. CRMsci http://www.ics.forth.gr/isl/index_main.php?l=e&c=663
4. Ronzino, P., Amico, N., Felicetti, A., Niccolucci, F., European standards for the documentation of historic buildings and their relationship with CIDOC-CRM, International Conference on Theory and Practice of Digital Libraries (TPDL 2013), 2013.
5. http://www.willpowerinfo.co.uk/glossary.htm
6. SKOS Core Guide http://www.w3.org/2004/02/skos/core/guide/
7. STELLAR project http://hypermedia.research.glam.ac.uk/kos/stellar/
8. TemaTres website http://www.vocabularyserver.com/
9. ISO 25964 thesaurus schemas. http://www.niso.org/schemas/iso25964
10. International Standard Organization (ISO). Documentation - Thesauri and interoperability with other vocabularies: Part 1, Thesauri for information retrieval, 2011. Report ISO 25964.
11. Alexiev, V., Isaac, A., On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI), International Journal on Digital Libraries, pp 1-10, 2015.
12. Cardillo, E., Folino, A., Trunfio, R., Towards the reuse of standardized thesauri into ontologies, Workshop on Ontology and Semantic Web Patterns (WOP2014), co-located with the 13th International Semantic Web Conference (ISWC2014), Riva del Garda, 2014.
13. SKOS Core Guide, W3C Working Draft 10 May 2005 http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/
14. SKOS-Reference http://www.w3.org/TR/2009/REC-skos-reference-20090818/
15. Getty website http://vocab.getty.edu/http://vocab.getty.edu/

Dati.CulturaItalia: a Use Case of Publishing Linked Open Data Based on CIDOC-CRM

Sara Di Giorgio[1], Achille Felicetti[2], Patrizia Martini[1], and Emilia Masci[3]

[1]Central Institute for the Union Catalogue of Italian Libraries (ICCU)
of the Italian Ministry of cultural heritage, activities and Tourism
(MiBACT), Rome, Italy
[2]PIN, VAST-LAB, Prato, Italy
[3] MIUR, Italy
{sara.digiorgio,patrizia.martini}@beniculturali
{achille.felicetti}@pin.unifi.it
{emilia.masci}@gmail.com

Abstract. In this paper we describe the pilot project dati.culturaitalia.it, which started in 2012 to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the web-portal CulturaItalia [1], the Italian national aggregator for Europeana [2]. CIDOC-CRM Ontology was used for transformation and representation of data widely pertaining to the cultural domain. RDF triples mapped into Erlangen CRM were then enriched with links to URIs identifying instances of internationally established RDF resources for geographic names, and instances of authority files for personal and corporate names, such as GeoNames and Virtual International Authority File (VIAF). CulturaItalia is the Portal of Italian Culture, promoted by the Italian Ministry of cultural heritage, activities and tourism (MiBACT), in which cultural institutions from all sectors and levels (national, regional and local) are involved. CulturaItalia also plays an important role for the development of Europeana, making available cooperative networks and agreements and coordinating technical activities leading to the establishment of Europeana environment.

Keywords: System interoperability, Data integration, Cross domain portal, CIDOC-CRM, Metadata Crosswalks, Linked Open Data, Semantic Web

1 Introduction

CulturaItalia [1], is the portal of the Italian Culture on-line since April 2008, managed by the Italian Ministry of cultural heritage, activities and tourism (MiBACT) through the Central Institute for the Union Catalogue of Italian Libraries (ICCU) [2]. The Web-portal indexes the main cultural databases and

[1] http://www.culturaitalia.it
[2] http://www.europeana.eu

gathers the metadata to Europeana, the public digital library promoted by the European Community. CulturaItalia is targeted to general users, by offering them a service for retrieving information on Italian culture from one access-point, and to more expert users, such as the operators in the cultural field, who can take advantage of a high-quality showcase to promote their own digital resources. CulturaItalia makes the digital resources interoperability possible, through a cross-domain Application Profile (PICO AP: PICO is the acronym for "Portale della Cultura Italiana On-line"), based on the Dublin Core Metadata Initiative technical guidelines. The Portal gives access to a rich "metadata" collection, which gathers and organizes information arriving from the various providers participating in the project. Users can discover different kinds of digital resources, describing the country's extensive cultural heritage (museums, photographs, libraries, archives, galleries, exhibitions, monuments, audio-visual works, etc.). The pilot project dati.culturaitalia.it started in 2012 with the aim to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the Web-portal. The application was designed by the CulturaItalia team with the technical and scientific support of Scuola Normale Superiore, and was developed by Meta s.r.l., to allow the resources aggregated by CulturaItalia to be involved into large semantic networks after exposing, sharing and connecting data according to LOD principles. A first release of this service is available on-line since 2013 [3] as a section, or sub-portal, of CulturaItalia dedicated to LOD. It presently makes available as LOD the Thesaurus PICO, adopted by the portal for facilitating the browsing of a variety of resources in its domain, and a selection of metadata sets from the Portal. The CulturaItalia team has chosen CIDOC-CRM, in the implementation of Erlangen CRM/OWL, to foster the interoperability in the cultural heritage sector. In the perspective of a future integration with the bibliographic heritage of the Open Catalogue of the National Librarian System (OPAC SBN), managed by ICCU, the Institute implemented, in 2014, a mapping activity, with the support of a team from VAST-LAB (PIN), to convert resources from OPAC SBN, encoded in UNIMARC format, in FRBRoo, adopting the CIDOC-CRM model.

2 CulturaItalia Application Profile and Thesaurus

CulturaItalia manages a catalog - called Index - which gathers and indexes metadata provided by the partners. The original data remain on the Web-site of the provider, to which the final user is redirected by CulturaItalia, through links, thus allowing to retrieve the original and complete information. For example, in the case of a photograph, in the CulturaItalia Index the preview image (thumbnail) is visible, together with some identifying data, and a link to the provider's website allows the user to visualize the photograph in its original format, accompanied by the complete information and services, in order to get the full benefit of the item. The resources in the Index are classified on the basis of the PICO Thesaurus, designed to manage and organize heterogeneous information, from

[3] http://dati.culturaitalia.it/

different cataloging systems. Browsing the Index, the user consults the metadata through a hierarchical classification of terms (facets). CulturaItalia is an "open" system: it grows up and develops together with the continuous enrichment of its metadata Index, through the metadata harvesting according to OAI-PMH, a protocol which allows the harvesting of metadata from content providers to one or more harvesters, adding services as indexing system or automatic classification. The Portal harvests metadata from different repositories and exports metadata to other national and international portals and repositories. At present CulturaItalia aggregates over 3 million metadata from 32 public and private partners including thematic aggregators, such as Internet Culturale, the portal of Italian Libraries, also created and managed by ICCU. Internet Culturale plays a key-role in guiding the libraries in the production of standardized digital cultural resources and metadata, according to the Italian standards. Metadata published in Internet Culturale are automatically transferred to CulturaItalia, and then, if the providing libraries agree, to Europeana.

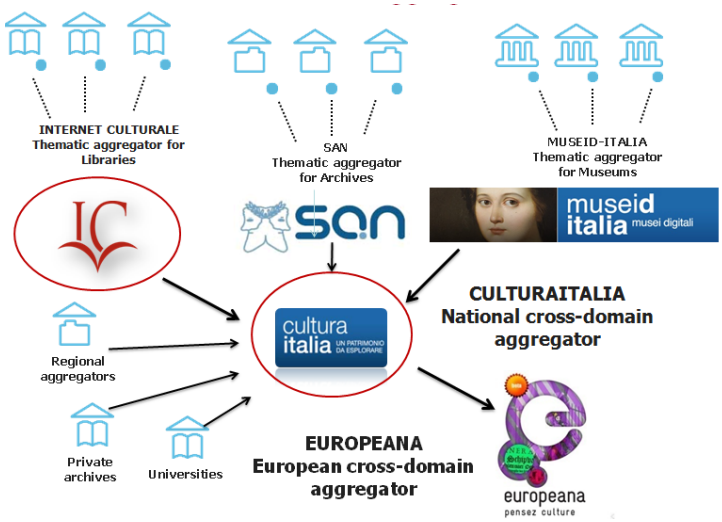


Fig. 1. CulturaItalia aggregation workflow

All content from external data sources are integrated in CulturaItalia in the form of metadata, thanks to the OAI-PMH harvesting protocol and published in the portal using a specific Application Profile (PICO AP), based on the international standard language, Dublin Core, that can describe, in a single scheme, every type of cultural resource, both physical and digital. Following DCMI recommendations, Scuola Normale Superiore di Pisa (SNS), which supports the ICCU working group engaged in the development of CulturaItalia, defined an application profile which joins DC Element set, Qualified DC terms and some further refinements and encoding schemes conceived for the applica-

tion of CulturaItalia. The PICO AP combines in one metadata schema all DC Elements, all DC Element Refinements and Encoding Schemes from the Qualified DC, and other refinements and encoding schemes specifically conceived to retrieve information pertaining to Italian culture. This Application Profile could be further expanded for harvesting possible unexpected contents in the future, by adding Refinements and Encoding Schemes that could be suitable for data retrieval. The PICO AP can be consulted at: http://purl.org/pico/picoap1.0.xml. Schemas used for the PICO AP are published on a PURL, under the domain PICO: `http://purl.org/pico/1.1/pico.xsd` and `http://purl.org/pico/1.1/picotype.xsd`. One of the most relevant encoding schemes introduced in the PICO AP is a Thesaurus specially conceived for the project itself, which comprehends hierarchically structured keywords indicating the topic of all the resources included into CulturaItalia (PICO Thesaurus 4.3). This ontology is also used to support the browsing into the Index of resources of Culturaitalia, therefore the assignment of a value taken from the PICO Thesaurus is mandatory for each metadata record. During the metadata generation, this assignment can be created for a whole repository or for a whole set, while in some other cases it was necessary to interpret a given value of the original database in order to create a mapping into the Thesaurus. The PICO Thesaurus is organized in four main categories: "Who" includes both people and corporate bodies; "What" comprehends tangible and intangible heritage, and all digital objects; "Where" covers Italian places (from regions to towns and villages) and "When" includes a list of chronological keywords associated to a sharp range of years. In order to be more compliant with international best practices, it seemed useful to adopt a SKOS format for the PICO Thesaurus. The SKOS format for Thesaurus PICO has also been designed to be extended and/or integrated with different thesauri pertaining to specific domains, managed by institutions that have a role in standardization, such as ICCD and ICCU, or to support multilingualism through the mapping between different national KOS.

3 Mapping between PICO Application Profile and CIDOC Conceptual Reference Model

PICO AP is a Dublin Core Application Profile. As already pointed out in the literature related to mapping between Dublin Core and CIDOC-CRM, for every value of the DC element "Type", specifying a type of a described resource, it must be specified a different mapping to a main entity of CIDOC CRM [3].

E.g.: IF DCMITipe = Image, THEN the described resource must be mapped as CIDOC-CRM entity = E38 Image. Consequently, each record encoded according to PICO AP will produce one main CIDOC-CRM corresponding entity, and the mapping of all the other PICO AP elements describing the resource will depend on the high-level mapping between the type of the resource and the corresponding CIDOC-CRM entity. Within PICO AP, dc:type element is mandatory and repeatable (occurrence: min 1, max unbounded) [4]. As a condition, it should always contain at least one value from DCMIType Vocabulary [5], (Col-

lection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text) or from PICOType Vocabulary [6] (CorporateBody, PhysicalPerson, Project). In the case that one PICO record contains more than one DCMIType and/or PICOType term, the mapping document (main DCMI/PICO Type term) specifies which must be considered the main term, according to which the mapping must be defined. Those simulations of complex mapping cases (between a PICO record containing more than one DCMI/PICO Type term and one CIDOC-CRM corresponding element) are described and are formulated on the basis of some real cases that can be found among CulturaItalia metadata resources (more than on a logic basis). Moreover, many of those cases are not real, and have been entered for completeness, just in case that in the future similar cases could occur. On the basis of the digital resources currently aggregated within CulturaItalia and of the PICO AP domain, the term DCMI Type = Physical Object is mapped to CIDOC-CRM entity = E22 Man Made Object (and not to E19 Physical Object). When DCMIType = Collection, the record generally contains many other DCMI/PICO Type terms. In all the cases, when "Collection" is present as a DCMIType, the PICO resource will always be mapped to CIDOC-CRM "E78 Collection" entity. On the basis of the mapping between the terms of DCMI and PICO Type Vocabularies, and CIDOC CRM entities, CulturaItalia resources encoded according to PICO AP can correspond to the following 12 CRM Entities:

E5 - - - - - - - Event
E22 - - - - - - Man-Made Object
E78 - - - - - - Collection
E28 - - - - - - Conceptual Object
E73 - - - - - - Information Object
E29 - - - - - - Design or Procedure
E33 - - - - - - Linguistic Object
E36 - - - - - - Visual Item
E38 - - - - - - Image
E39 - - - - - - Actor
E40 - - - - - - Legal Body
E21 - - - - - - Person

From this high-level mapping, based on the type of the described resource, derive different mappings between the various types of PICO AP resources and a corresponding CIDOC-CRM entity.

For a PICO resource with DCMIType= "PhysicalObject" (= crm:E22 Man Made Object), the PICO AP element $< pico : author >$ must be mapped as shown in figure 2.

For a PICO resource with DCMIType= "StillImage" (= crm:E38 Image), the same PICO AP element $< pico : author >$ will be mapped as shown in figure 3.

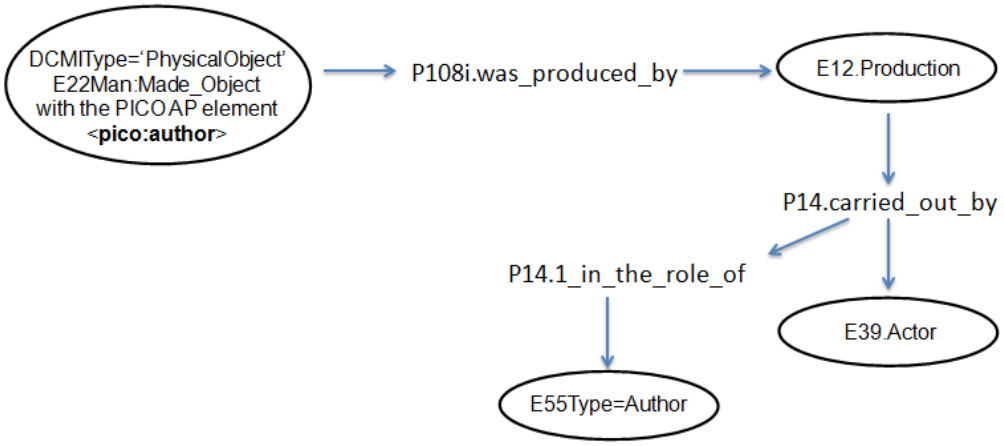


Fig. 2. Mapping of the PICO AP element $< pico : author >$ for a PICO resource with DCMIType= "PhysicalObject"

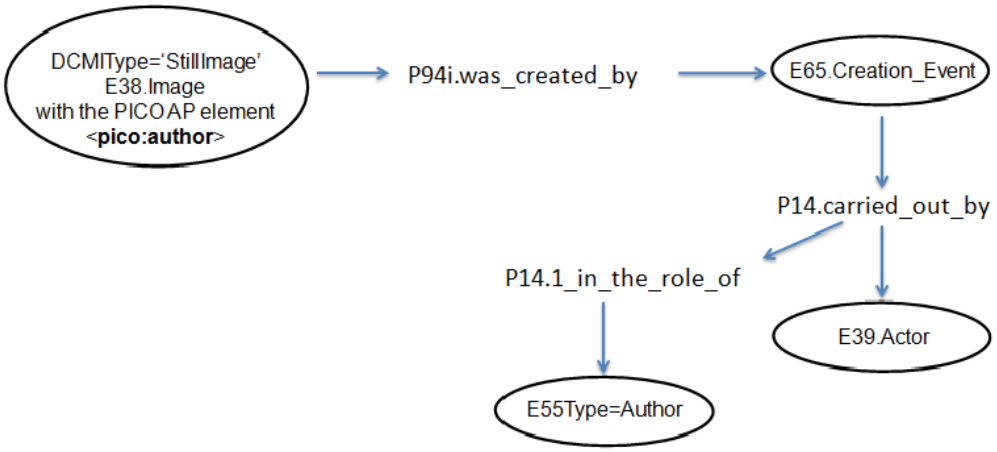


Fig. 3. Mapping of the PICO AP element $< pico : author >$ for a PICO resource with DCMIType= "StillImage"

As CRM entities are ordered in a poly-hierarchy, and as the properties associated to each class are inherited by the subclasses, it is possible to group the 12 identified entities (and, consequently, the mapping to be implemented) into 4 main groups:

1. E18 Physical Thing contains the mappings for: E22 Man Made Object and E78 Collection
2. E28 Conceptual Object contains the mappings for: E73 Information Object, E29 Design or Procedure, E33 Linguistic Object, E36 Visual Item, E38 Image
3. E39 Actor contains the mappings for: E40 Legal Body, E21 Person
4. E5 Event

The detailed mapping containing four mapping tables (one for each of the above-listed CRM main entities) is available on-line within the document: "Mapping between PICO Application profile and CIDOC Conceptual Reference Model" [7].

Figure 4 presents the main elements of the mapping related to E18 Physical Thing that contains the rules for E22 Man Made Object and E78 Collection.

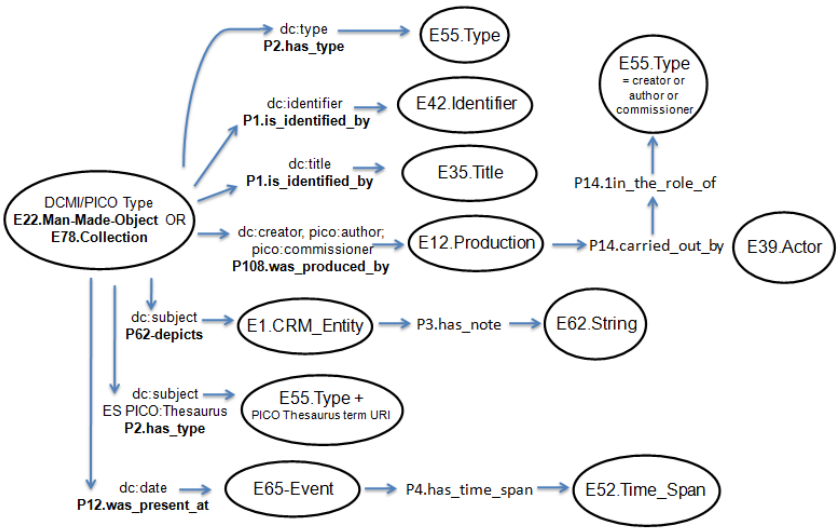


Fig. 4. Mapping of E18 Physical Thing

4 Dati.CulturaItalia

The pilot project dati.culturaitalia.it started in 2012 to build up a Linked Open Data (LOD) Service that will progressively make available open datasets from the web-portal CulturaItalia. A first release of this service is available on-line

since March 2013 [8] as a section of CulturaItalia dedicated to LOD. It presently makes available as LOD the Thesaurus PICO and metadata aggregated by the portal and licensed under CC0 1.0 - Universal Public Domain Dedication. These are data coming from: Accademia di Santa Cecelia, ArtPast Project, Digibess, ICCU, Internet Culturale, Michael Italia, Polo Museale Fiorentino, Regione Marche and Anagrafe delle Biblioteche Italiane. More datasets will be increasingly published as LOD, as soon as they will be delivered under CC0. CulturaItalia platform extracts the datasets, encoded in XML PICO format, that have been submitted by providers agreeing to take part of the pilot and to convert the PICO metadata into CIDOC [9] standard, according to the mapping document elaborated by M. E. Masci (Pisa, SNS). The mapping is implemented in an XML stylesheet and the result is an RDF/XML representation of each data provider's metadata. Then the CulturaItalia repository allows for the semantic enrichment with four types of reference resources (authority files):

- VIAF (Virtual International Authority File: www.viaf.org)
- GeoNames (www.geonames.org/)
- PICO Thesaurus in SKOS
- DCMI Type vocabulary

The SPARQL endpoint provides access to RDF metadata structured according to the CIDOC - Conceptual Reference Model in the implementation of Erlangen CRM/OWL. Data can be searched over three querying interfaces, corresponding to three sections of dati.culturaitalia.it:

- Text search: here it is possible to perform free text searches over all triples contained in dati.culturaitalia.it.
- SPARQL query: here you can try your hand at a SPARQL query. There are also some examples of queries.
- iSPARQL query: here there is an even more complex querying interface for advanced users.

Dati.culturaItalia.it exposes an OAI Provider that makes available XML or RDF metadata structured according to different schemas:

- oai-dc (xml): OAI-PMH schema adopted by Open Archives Initiative Protocol for Metadata Harvesting
- pico (xml): PICO Application Profile, the CulturaItalia Application Profile
- edm (rdf): Europeana Data Model, adopted by the portal Europeana EDM [10]
- cidoc (rdf): CIDOC - Conceptual Reference Model in the implementation of Erlangen CRM / OWL

5 Mapping between UNIMARC Bibliographic Format / SBN MARC and FRBRoo and next steps

ICCU is moving another step towards the Italian Linked Cultural Data Cloud by starting the mapping study of data from the OPAC SBN (On line Public

Access Catalog of National Library Service) in UNIMARC format to the class and the properties of FRBRoo, on the base of the model CIDOC CRM.
The collective catalogue of National Library Service provides access to 13.759.767 bibliographic records that contains:

- descriptions of documents acquired from SBN libraries starting from the '90s or since single libraries entered the SBN
- descriptions "book in hand" of documents of XVI - XX centuries
- descriptions obtained from catalogues on paper previous to 1990

In 2014 a working group formed by experts from ICCU and VAST-LAB (PIN) was established with the objective to analyze and test the publication of a subset of significant data in UNIMARC format as LOD according to the document *FRBR object-oriented definition and mapping to FRBR-ER (version 0.9)*. In particular, this activity focused on:

- analyzing and defining a basic methodology for creating Linked Open Data from bibliographic archives according to the international standards and cataloguing rules adopted by SBN;
- designing a schema with the conceptual description of how to relate SBN bibliographic information in a semantic way. FRBRoo, an harmonization between FRBR original conceptual model and CIDOC CRM, has been chosen as the reference intellectual guide for this activity
- selecting a first set of bibliographic records to be exported from OPAC SBN in UNIMARC format
- defining all the required namespaces and URI mechanisms to create meaningful identifiers for the converted UNIMARC entities

Activities performed by the working team lead to the definition of a mapping document describing the conceptual mapping between UNIMARC fields and FRBRoo entities and properties, with specific definition of mapping paths for every possible combination or special use cases of UNIMARC encoded information available in the SBN archive. The selected records subset was used for testing the conceptual coherence of the model in order to identify possible conflicts and to fix co-reference and cross-reference issues that might have arose.

Specific exporting scripts have been developed to encode the UNIMARC bibliographic information in a standard RDF format, to transform it in a machine-readable version using a formal language. Bibliographic information created in this way was afterward enriched with entities coming from VIAF, GeoNames, Linked Heritage, DBPedia, and other available online Linked Open Data resources.

A web tool has also been created to store semantic records and to query and retrieve relevant bibliographic data according with given semantic criteria. The tool is composed of various modules efficiently interacting with each other and based on open source technology. The modules include:

– an online triple store based on Sesame to accommodate the RDF triples created by the exporting framework and to manage the complex network of relationships defined by means of it;
– a set of responsive web interfaces based on Ajax/JQuery technologies and implementing the various features of semantic query and presentation of the relevant results. A basic faceted system for a more efficient browsing of the results was also implemented within the same interfaces.

The web tool also offers the possibility to download the full Linked Open Data network of bibliographic information in an RDF compatible format for local use. Further work on this topic will necessarily require a data cleaning phase for consolidating the legacy database in order to create a better representation of its content during the mapping and conversion process. Additional activities will concern the creation of a SPARQL end point for advanced semantic queries, the improvement of the web interface to allow connection of various libraries to the SBN index, to facilities retrieving and FRBRoo encoded triples representing entities of interest (work, expression, etc.) and the export of the same information in a standard Linked Open Data format for them to be used by other bibliographic tools and in other similar contexts. Data validation cycles to ensure the full compatibility of the formats with the fundamental principles of Linked Open Data and Semantic Web philosophy will be also performed, as well as multiple tests on the internal coherence of the newly created dataset.

References

1. http://www.culturaitalia.it/
2. http://www.iccu.sbn.it/opencms/opencms/it/
3. Main references for the present mapping: M. Doerr, Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM, Technical Report 274, ICS-FORTH, July 2000: http://www.cidoccrm.org/docs/dc_to_crm_mapping.pdf; K. Kakali, M. Doerr, C. Papatheodorou, T. Stasinopoulou, DC.type mapping to CIDOC/CRM, DELOS WP5-Task5.5, Department of Archives and Library Science / Ionian University, 26/01/2007: http://www.cidoc-crm.org/docs/WP5-T5_5-DC2CRMmapping-060728v0_2-final.doc ; I. Lourdi, C. Papatheodorou, M. Doerr, Semantic Integration of Collection Description. Combining CIDOC/CRM and Dublin Core Collections Application Profile, D-Lib magazine, July/August 2009, vol. 15 n. 7/8, ISSN: 1082-9873: http://www.dlib.org/dlib/july09/papatheodorou/07papatheodorou.html; M. Doerr, Updated graphical representation of the harmonized EDM-CRM-FRBRoo-DC-ORE models, September 2011: http://www.cidoc-crm.org/docs/EDM-DC-ORE-CRM-FRBR_Integration_ORE_fix.ppt
4. PICOAP/dc:type: http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml#type
5. DCMI Type Vocabulary: http://dublincore.org/documents/dcmi-type-vocabulary/

6. PICO Type Vocabulary: http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml#PICOType
7. M. E. Masci, Mapping between PICO Application Profile and CIDOC Conceptual reference Model version 1.0, 2013-01-24
8. See at http://dati.culturaitalia.it
9. See the document at http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/mapping/pico_cidoc/mapping_PICO_CIDOC-CRM_ITA-ENG.pdf
10. See the document at http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/mapping/pico_edm/Mapping-PICO-EDM-2.0.pdf

CIDOC CRM and Epigraphy: a Hermeneutic Challenge

Achille Felicetti[1], Francesca Murano[2], Paola Ronzino[1], and Franco Niccolucci[1]

[1] PIN, VAST-LAB, Prato, Italy
[2] Università degli Studi di Firenze, Italy
{achille.felicetti,paola.ronzino,franco.niccolucci}@pin.unifi.it
francesca.murano@unifi.it

Abstract. This paper identifies the main concepts involved in the study of epigraphy and proposes the use of CIDOC CRM to encode epigraphic concepts and to model the scientific process of investigation related to the study of epigraphy. After analysing the existing CIDOC CRM entities and those provided by the CRMsci and CRMarchaeo extensions, we introduce more specific epigraphic classes to be used as the basis for creating a new extension, CRMepi, which is more responsive to the specific needs of this discipline.

Keywords: CIDOC-CRM extension, Epigraphy, EpiDoc, CRMepi

1 Introduction

Material sources bequeathed to us from antiquity represent an immense treasury of knowledge of a lost world. In addition to literary sources, numismatics, epigraphy and archaeology provide important information we would not otherwise possess, and often the objects they document constitute the only evidence we have of an ancient population. In particular, among these sources, inscriptions are particularly essential because they represent the most direct and resonant voice of the ancient people that have handed them down to us, as deathless words in stone. Generally, the preserved inscriptions were originally conceived to endure over time and to be transmitted to the future. For these reasons, these sources need to be digitised and integrated in some way, along with other cultural heritage information. Inscriptions are complex objects, and their study requires careful analysis from different points of view.

What characterises this class of objects is that they form a whole with their physical support. Indeed the meaning of an epigraph cannot be fully understood without the analysis of the object or monument or other archaeological object on which it appears, just as one cannot fully understand the nature of that particular archaeological object without thoroughly investigating the sense of the inscription or iconographic representation it hosts.

Furthermore, the inscription itself is, from a conceptual point of view, an element with physical characteristics that are themselves bearers of meaning and of valuable information going far beyond the inherent meaning of the text. The

shape of the letters, their spacing, the direction, technique and other similar characteristics provide precious clues to the times, makers and functions of the inscription in question.

The tools available today seem to be insufficiently flexible and efficient to implement a comprehensive and useful level of digitisation. Relational databases, for instance, are not fully functional for this goal because of the rigid structure they provide, which is quite inadequate to describe textual entities and their possible annotations. Databases are even less suitable to describe the complex web of relationships that links archaeological objects with the inscriptions they bear and their meaning. Since the beginning of the studies on digitisation of texts and other unstructured data, the XML family of technologies emerged as the tool that could free the information from the rigid structures of relational databases built around tables and records. Epigraphy has also largely benefitted from XML. It is sufficient here to say that EpiDoc [1], the standard currently used to encode epigraphic information in electronic format, is also based on XML, the guiding principles of which it is both the bearer and the beneficiary. However, in the modern digital world, where the imperatives are those of interoperability and integration, the use of more efficient tools such as ontologies and conceptual models seems to be of crucial importance. In this paper, after investigating the problem, identifying the key concepts involved and giving an overview of the existing solutions, we will try to give a coherent description of the new possibilities offered by semantic tools to deal with epigraphic entities. In particular, we will use the intellectual model of CIDOC CRM and its extensions (CRMsci and CRMarchaeo) to provide them with shape and consistency, and to try to sketch a new CIDOC CRM extension (CRMepi) to be used in epigraphic studies.

2 What is Epigraphy?

2.1 Epigraphs and Epigraphy

Although epigraphy is a scientific discipline with a centuries-old tradition, a single and fully accepted definition of its object, the epigraph (or inscription), has not yet been formulated. The definition most commonly found in manuals of the discipline defines the inscription more or less as a direct evidence of the past inscribed on stone or other durable materials. Guarducci, in particular, puts emphasis on the materiality of the support, which, among other things, makes it possible to distinguish the object of epigraphy, as opposed to papyrology. She identifies what epigraphy is concerned with, as opposed to papyrology, and she identifies two particular characteristics of epigraphic documentation that differentiate it from other historical primary sources: uniqueness and authenticity [2]. More recently, Panciera has proposed to define an epigraph as “any particular type of written human communication of the sort that we would today call unidirectional, in the sense that it does not anticipate that a response will be provided to the sender, and which has the characteristic of not being addressed

to a person or to a group but to a collectivity, and which for this reason is made with the location, writing technique, graphic form and impagination, mode and register of expression chosen because they are most suitable to the attainment of its intended goal, and which differentiates itself in this manner from other forms of contemporary verbal communication (oral, literary, or documentary)" [3]. This definition is undoubtedly more complete than the previous ones, but is still not sufficiently exhaustive with regard to the huge variety of documents that are the object of epigraphic studies and may be included in epigraphic *corpora*.

The problem originates from the plethora of tools, techniques and purposes that can be employed in order to characterise an inscription. These constitute a complex and intertwined series of elements that do not, however, suggest a comprehensive definition. We must also consider that different cultures have given rise to different traditions of writing and thus also of epigraphy. Consequently, a consistent and comprehensive definition of the term "inscription" should consider not only epigraphic products from the major Western traditions (Latin and Greek), but also those relevant to "minor" traditions (e.g., the languages of ancient Italy) and non-Western ones (e.g., the heterogeneous legacy of the Semitic world). In addition, we should consider that there is an ambiguity of reference with respect to the term "epigraph", which can be used to indicate both the plain text and the combination of text and physical support, especially in those cases in which the support has been created expressly for bearing the text; in fact, inscriptions may also appear on artefacts made for different purposes (such as vessels) or even on non-artefacts or on natural surfaces (caves, cliffs, etc.). It is obviously not up to us to say the last word about this topic, or to define what epigraphy is and what distinguishes it, for example, from papyrology and other textual studies. Nevertheless, it is certainly very difficult to find a comprehensive ontological definition for classes of objects that have been from time to time assigned to one category or another, mainly depending on the discipline involved in their study.

2.2 The nature of an epigraph

From a logical point of view, and in accordance with the tradition of epigraphic studies, an inscription can be analysed according to three main aspects: the text-bearing object or monument (obviously involving archaeological topics), the text (and the obvious correlations with content and linguistic aspects *lato sensu*) and the feature engraved on the support in the form of letters or other symbols, which is the central element that characterizes and differentiates an epigraph from any other manifestation of written communication. Whereas in documents on papyrus or parchment (*et similia*), the materiality of the features is not relevant in comparison with the morphological and typological characteristics of the handwriting, investigated by palaeography, they are essential in inscriptions and their analysis precedes and prepares palaeographic studies. Such features represent the peculiar object of epigraphic studies, and analysis of them is as fundamental as that of the archaeological, palaeographic, linguistic

and historical aspects. Paramount importance must be ascribed to the communicative purpose of the epigraphic text, since this is exactly what distinguishes an inscription from any other feature of any kind that may be present on a given support; in other words, given the intentionality of a feature, a figurative decoration is distinguished from an inscription by both semiotic and linguistic purposes (the explicit will of communicating a message), on which a different communicative-informational structure also depends. According to a semiotic analysis, an intentional feature (i.e., one voluntarily created by man to convey a message) occurring on a given support can appear in the following forms (see [4] [5] [6]):

- Features not belonging to any writing system, i.e. the figurative decoration, even when it has value of icon or symbol (e.g., the sign of the Christian cross)

- Features belonging to 'non-linguistic' writing systems, i.e. signs of pure semasiographic systems of writing, 'language-independent', used to represent concepts and not related per se to a given linguistic structure

- Features belonging to a linguistic writing system, but not used per se: in this case we are in the presence of signs with a linguistic value but written for a purely decorative purpose or used as symbols (e.g., the A and Ω signs used as symbols of the beginning and end in the Christian tradition)

- Features belonging to a linguistic writing system and used per se, i.e. signs of a glottographic writing system, (variously) depending from a given linguistic system, thus taking on a real linguistic value

We can talk of inscriptions in the latter three cases, i.e., when we recognise signs belonging to certain writing systems; nevertheless, we can talk about written communication only in the presence of signs used *per se* as encoders of linguistic signs (and structures), and therefore, of a text.

Finally, we can summarize as follows:

- not-glottographic feature, a figurative decoration, but also a sign of pure semasiographic systems of writing

- glottographic feature, not necessary codifying a linguistic expression, since a sign can be used with different purposes than a linguistic one

3 Standards for (Digital) Epigraphy

The edition of ancient texts boasts one of the earliest and more consistent systems of standardisation in the field of Humanities: the Leiden Conventions [7].

This standard, which arose from the need to publish texts using a shared notation to describe the various observable phenomena they show, was created by an international group of scholars gathered in Leiden in 1931 and is the standard still adopted in modern epigraphy. Many of the well-established and growing database-based epigraphic *corpora*, including the Epigraphische Datenbank Heidelberg, the inscriptions section of the Deutsches Archäologisches Institut and the Epigraphic Database of Rome, also provide an extensive text field containing the text of the inscription in Leiden format, besides the typical descriptive fields used for metadata, such as find location, date, dimension and so on. The Leiden Conventions specify how features of an inscription besides the text itself should be represented in print, by using a set of standard symbols and text decorations to reproduce the state of the original document and to report the editors' interpretations.

However, with the advent of the digital era, epigraphists had to face a set of problems very similar to the ones brilliantly solved by the Leiden Conventions. An electronic format that could allow digital publishing, storage and exchange of epigraphic information in a consistent and shared format was needed. From this need arose EpiDoc, a collaborative format designed to transcode in digital format the Leiden-encoded printed editions. The initiative began in the 1990s in response to the request for a free and unrestricted set of tools supporting the creation of online epigraphic archives, which was expressed during the same period in the course of a series of conferences on epigraphy and IT. The XML format was identified as being the most suitable for this purpose and the first EpiDoc DTD was released and quickly adopted by a relatively wide community of researchers. Basically, EpiDoc is an application profile of TEI specifically adapted to the needs of epigraphy. This profile has extensively evolved from its first draft. As of today, EpiDoc provides features for the recording of the materiality and history of text-bearing objects, as well as features for scholarly editions of the text, such as commentary, illustrations, bibliography, and publication data. EpiDoc also offers facilities for the detailed description and editorial representation of the texts themselves, including transcription in the technical sense of reporting readings and representing the writing system, form, appearance, layout and editorial interventions in the text.

The EpiDoc system, despite its undoubted merits, still presents some issues, especially with respect to the inline text encoding features, arising from the fact that there are no native tools fully able to support the EpiDoc format for sessions of text editing, and thus to simplify the encoding operations. Essentially, EpiDoc-based text edition still remains a manual task, which greatly complicates the digitisation of large *corpora* of inscriptions. From a technological point of view, the choice of reproducing the paper publication format by means of XML mark-ups also raises an issue related to the style sheet necessary for the optimal rendering of the XML encoded text and therefore inseparable from it. This could represent a further portability issue for the sharing of the information from one archive to another and on the Web. EpiDoc is also unable to guarantee the typical "relational" features offered by a database, since it is lacking in

all the paraphernalia necessary to describe the complex web of relationships that characterize the various aspects of epigraphy. Only ontologies and similar semantic tools seem to be able to merge the advantages and flexibility typical of XML with the characteristic "relationality" of databases.

4 A tentative CIDOC CRM representation

4.1 Defining the concepts

In past years, some attempts have been made to use CIDOC CRM for the description of epigraphic entities. One of the first such projects was VBI-ERAT-LVPA [8], the aim of which was to use CIDOC CRM for the integration of epigraphic digital archives. This project had the merit of having given a first reply to the question of how to describe an epigraph and its various components using conceptual tools, but it did not provide definitive conclusions on the subject.

More recently, some methodological proposals have been put forward to combine EpiDoc and CIDOC CRM features and to harmonize the features they provide [9]. In particular, the EAGLE project [10], which aims to create a portal for the integration of some of the most important existing epigraphic archives, is currently engaged in the mapping between EpiDoc and CIDOC CRM, a task of great interest that will surely provide excellent suggestions for the definition of a possible extension and for the convergence of the two models. The ARIADNE project [11], although its main focus is the integration of archaeological archives, is also involved in the study of inscriptions, not only as archaeological objects but also as regards interoperability between archaeology and epigraphy. During several workshops and summer schools, the issue has been extensively discussed and outlined in its main aspects. This paper is also one of the results of these activities.

As mentioned, the purpose of this work is to lay the groundwork for a possible epigraphic extension of CIDOC CRM (CRMepi) through the conceptual analysis of the specific entities and problems with which epigraphy is concerned. We are aware that the lack of a comprehensive conceptual definition of the identity of an epigraph obviously complicates the formal definition of epigraphic entities using tools provided by ontologies and conceptual models at our disposal. Some general observations may however be made and the conceptual model of the CIDOC CRM can certainly be used as an intellectual guide to reasoning on the concepts involved in this process (see Figure 1).

4.2 The physical support

In the case of epigraphy, an essential element that emerges from the above discussion is the close relationship that tightly binds the inscription with its physical support. This close cohesion between the support and the text, as already mentioned, is, for instance, what distinguishes an epigraph from a papyrus, the study of which mainly concerns textual analysis. For some inscriptions, however, the

shape, the materials, the production techniques and all the attributes of the physical object that hosts the epigraph can become fundamental not only for their understanding but also for the definition of their nature. If we focus on the support on which the inscription was engraved, we note that CIDOC CRM offers plenty of concepts with which we could describe it. The physical support is in fact very often an archaeological object, a class of objects which has frequently been investigated in a CIDOC CRM perspective. In terms of integration and interoperability it is also important to note that, thanks to its nature, the support constitutes one of the main points of contact between epigraphy and archaeology. The specific archaeological aspects (discovery, provenance, archaeological context etc.) relating to the physical support can be documented using the CRMarchaeo extension [12].

In relation to epigraphy, it should be noted that very often the physical support has been designed and built specifically to accommodate the inscription. In this case the CIDOC CRM *E84 Information Carrier* entity seems definitely to be an optimal choice. However, this condition does not always occur: certain inscriptions may in fact even have been placed on objects not specifically designed to accommodate an inscription, as in the case of buildings, vessels or other objects of daily use on which the inscription may have been placed at a later time. In this case, the use of a more generic class, like *E22 Man-Made Object*, seems more appropriate. There are also cases in which the inscription is placed on natural surfaces not created by human activities, such as inscriptions on rocks, in caves or other similar natural places. The use of the superclass *E19 Physical Object* sufficiently broad so as to include every possible kind of physical support and would be in this case a more suitable choice. Each of these classes can still be linked with the physical features they bear, via the *P56 bears feature* property, having the *E19* class as domain and thus being inherited by all its subclasses. The EpiDoc elements used to mark archaeological information concerning physical objects or monuments (such as the *supportDesc*, *material*, *objectType* and *dimension* tags) can easily be mapped using these CIDOC CRM entities.

4.3 The inscription

CIDOC CRM provides a specific class to model the concept of inscription: *E34 Inscription*. The scope notes of this class state that "this class comprises recognisable, short texts attached to instances of *E24 Physical Man-Made Thing*". We need at this point to make sure that this class is consistent enough with the concept of inscription in the epigraphic sense, so as not to risk incurring conceptual ambiguities. Although many inscriptions bear short texts, the brevity or length of an inscription is not among its main characteristics. In fact, there are inscriptions occupying entire walls (the Gortyn Law Code or the *Res Gestae Divi Augusti*, for example) and in any case the "short text" of the *E34* class remains too vague and undefined for the purposes of our investigation. The *E34* class also belongs to the classes of conceptual objects which in turn are defined as "non-material products of our minds and other human produced data",

something that renders only in part the essence of an epigraph, not taking into any account its "materiality" which is a fundamental component of its identity. The study of epigraphy typically moves from the analysis of the physical characteristics of inscriptions before getting to their archaeological, palaeographic, linguistic and historical characteristics. In this sense, an inscription intended only as a conceptual object does not seem to capture fully the very nature of the epigraph itself. Thus, the etymology of the word "epigraph" indicates as a fundamental condition of its identity its being written on something. In all these ways it seems to present a much closer resemblance to the classes created for the description of physical features, and more specifically the *E25 Man-Made Feature*. We have managed to create some new and more appropriate classes to be used in documenting epigraphic concepts, and in particular:

- ***EPI1 Epigraph***. Subclass of *E25 Man-Made Feature* intended to describe a particular feature created by humans, in various ways and on various kinds of support, mostly rigid ones, with the declared purpose of conveying a specific message towards a given recipient or group of recipients.

- ***EPI2 Engraving***. Subclass of *E12 Production* indicating the activity of creating inscriptions in an epigraphic sense by using various techniques (painting, sculpture, graffiti etc.) and by means of specific tools on a given physical carrier. The definition of this activity allows us to make a better distinction between the creation of inscriptions and the production of the physical carriers that host them (two activities that are not always and not necessary contemporary), and to distinguish more accurately the creation of an epigraph from that of a story or poetry or other literary texts written, for example, on a papyrus.

- ***EPI3 Epigraphic Field***. Subclass of *E25 Man-Made Feature*. This represents another important element of epigraphy, usually understood as the surface or portion of the physical carrier reserved, delimited and arranged for the purpose of accommodating an inscription, to highlight it and to isolate it from the other parts of the object or building to which it belongs. There are various types of epigraphic fields, and among the most important of these epigraphists usually distinguish those created during the same production event of the carrier and those added to it at a later time. From a conceptual point of view, the epigraphic field is a feature designed to accommodate another feature (the inscription). EpiDoc also provides specific entities for the description of these elements (the tag *layoutDesc* for example) that can be easily mapped on this class. To define the relationship occurring between *EPI3 Epigraphic Field* and *EPI1Epigraph*, the new property *EPP2 is included within* has been proposed as a sub property of *P56 bears feature*, which is in turn a shortcut of the full path relating *E19 Physical Object* through *P59 has section (is located on or within), E53 Place, P53 has former or current location (is former or current location of)* with *E26 Physical Feature* as described in the related scope notes.

4.4 The text of the inscription

As already noted, conceptually an epigraph significantly differs from the definition of the class *E34 Inscription* of CIDOC CRM. To avoid semantic and linguistic ambiguities we decided to use its superclass *E33 Linguistic Object* to describe the text (intended as a linguistic production), which the inscription records. Before proceeding any further, it would be appropriate to clarify the relationship between the epigraph, intended as a feature consisting of a set of signs, and its text as obtained through observation and decoding of those signs and the interpretation of the linguistic signs they refer to.

One of the most important operations carried out by epigraphists for their study, and especially for publication, is the so-called "reading" of the epigraph, consisting of a deep and accurate analysis and study of the surface and the signs followed by establishing as faithful as possible what is shown by the physical feature. This scientific process can be modelled by means of the concept of observation documented in the scientific extension of CIDOC CRM (CRM-sci), and more precisely by means of the *S4 Observation* class. This class seems particularly appropriate to document this kind of scientific analysis. It could also avail itself of specific instrumentation to assist reading, such as microscopes or magnifying glasses, especially in case of inscriptions of reduced dimensions. It is also possible, thanks to this class, to document the different processes of analysis and study carried out over the years on a particular object by various scholars, and to report details on the tools and methods used. The observation (*S4*) class is used to define and identify (*O16 observed values*) the graphemes (i.e., the symbolic object *E90* used to encode, on an abstract level, the linguistic units in the text) of which the engraved signs on an *EPI1 Epigraph* are the concrete graphical manifestation; this relation is made explicit by means of the *P128 carries* property. The opportunity to instantiate a more specific subclass of *E90 Symbolic Object* in order to provide a better description of these ideal graphemes requires a more thorough discussion.

The graphemes inferred from the observation of the epigraph represent the level of the intellectual decoding and understanding of the signs and constitute the basis for the subsequent operations of transcription usually carried out by epigraphists, in particular for so-called *diplomatic* transcription (i.e., a specific transcription recording only the characters as they appear on the support, without any editorial intervention or interpretation), which is also of great importance from the point of view of publication. The publishers of an epigraph, for practical reasons, generally those of typeface, perform these transcriptions using Latin or Greek characters, even in case of non-Latin and non-Greek inscriptions (Etruscan inscriptions, for example). To document this process we have created some specific classes:

- ***EPI4 Transcription***. Subclass of *E7 Activity* describing the specific operation of transliteration that, starting from the symbols observed on the epigraph (*E90 Symbolic Object ->P16 was used for*), leads to the creation (*P94 created*) of a set of instances of *E73 Information Object* recording the

transcription(s) performed.

– ***EPI5 Writing System***. Subclass of *E29 Design or Procedure*, which refers to a specific sequence of characters (graphemes, *E90 Symbolic Object*) used both to write and to transliterate the text of the epigraph (e.g. Latin letters).

As mentioned, EpiDoc provides an entire section of tools for the encoding of the diplomatic transcriptions, implementing the Leiden Conventions in the form of specific inline XML tags designed to mark and describe each part of the text. An instance of the *E62 String* class, in association with one of the *E73 Information Object* (with type = *diplomatic*), might accommodate the EpiDoc-encoded text by means of a *P3 has note* property, to strengthen integration. The *E33 Linguistic Object* class, as we already mentioned, was used instead to represent the text of the inscription, i.e., the intended linguistic entity as resulting from an intellectual and linguistic process of creation, witnessed by the epigraph and also inferred from the various transcriptions (*E73 Information Object*). The latter relation is established via the *EPP2 has transcription* property, a sub property of *P130 shows features of*. The connection between *E33 Linguistic Object* and *EPI1 Epigraph* is instead represented by using the *P62 depicts* property, which perfectly renders the nature of this relationship. It should be noted that the graphemes (*E90*) also remain in constant and close connection with the *E33 Linguistic Object*. We have used the *P67 refers to* property to describe this connection. In future developments, however, we shall assess the question of whether it is opportune to create a new sub property that better expresses this type of link.

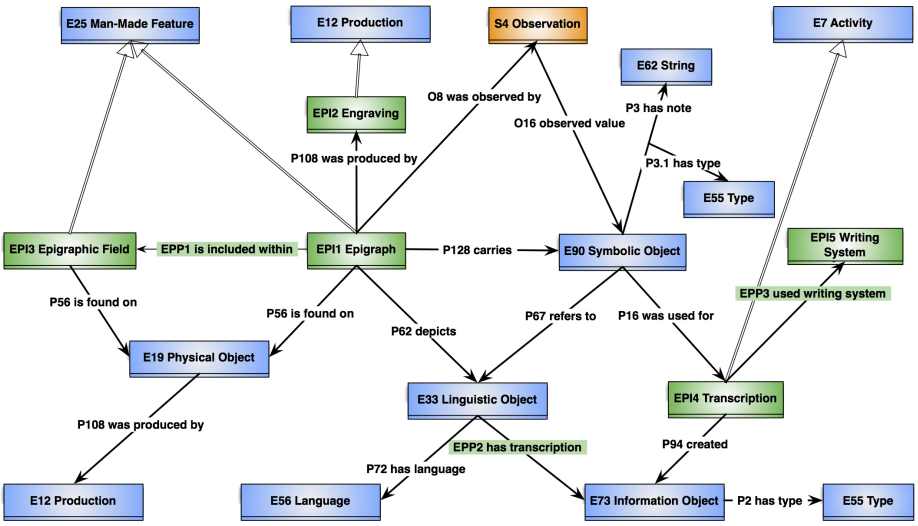


Fig. 1. The CIDOC CRM epigraphic model and extension

5 An example: the Oscan inscription VE 150

To test and demonstrate the potentialities of the proposed model, we chose an epigraph in the Oscan language from Pietrabbondante (Isernia, Italy), stored at the Archaeological Museum of Naples (Italy). It is a *cippus* (MANN.2525) bearing a dedicatory inscription (VE.150) [13] of an Italic sanctuary in the Oscan alphabet. Both the artefact and the inscription have been dated to around the second century BC. Many scholars have studied the inscription, but for the purposes of the example presented here we have taken into account the analysis and interpretation provided by Mommsen in 1850 [14] (see Figure 2).

The *cippus* was created specifically to host the epigraph. In this case, it seems more appropriate to use the *E84 Information Carrier* class rather than its *E22 Man-Made Object* superclass to encode it. The use of *EPI1 Epigraph* and the related production event (*EPI2 Engraving*) allows us to distinguish the event of creation of the epigraph from that of the archaeological object, although in this case the two events happened simultaneously, since the inscription was in fact sculpted (*P32 used general technique ->E55 Type = sculpture*) contextually to the production of the *cippus*. The definition of the same *E52 Time Span* for both the events indicates their contemporaneity.

The *E90 Symbolic Object* is represented by a set of characters of the Oscan alphabet evidenced through the property *P3.1 has type*, a sub property of *P3 has notes* usually used to record specific notes concerning peculiar aspects of a given entity such as the writing system used, as happens in this case. The alphabet to which the graphemes of the *E90 Symbolic Object* belong is an element of capital importance, especially in the study of non-Latin and non-Greek inscriptions, which needs to be expressed in a richer and more specific way. We will consider the definition of adequate classes and properties for modelling this concept in future revisions of the model. Through analysis of the inscription (*O4 Observation*), in 1850 Theodor Mommsen provided a reading and a transcription of it (*EPI4 Transcription ->P94 created ->E73 Information Object*) using the Latin alphabet (*EPP3 used writing system ->EPI5 Writing System*) to create both the diplomatic and interpretative transcriptions of the text. This same set of classes and properties can be instantiated several times in case of new or different readings, transcriptions and interpretations of the same epigraph by other scholars, in order to create a chain of events able to represent the history of studies of the object.

Diplomatic transcription constitutes the basis for the interpretative edition of the *E33 Linguistic Object*, i.e., the text intended as a linguistic production, encoded by means of a given writing system in the epigraph. The *E33 Linguistic Object* is therefore linked both to the *E90 Symbolic Object* (i.e. the Oscan graphemes, units of writing system on an abstract level) and to the *EPI1 Epigraph* (the concrete manifestation of such Oscan units as physical features), through the *P67 refers to* and the *P62 depicts* properties respectively. Since the *E33 Linguistic Object* is an expression of the Oscan language (*P72 has language*), it can be provided with a translation into any other language (*P73 has*

translation), for example, into English, in order to make the content more understandable.

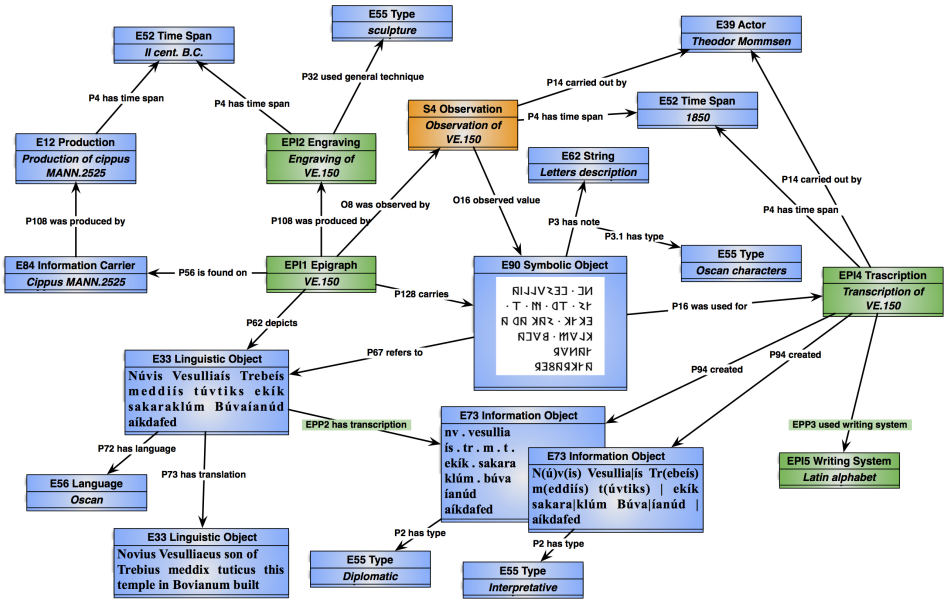


Fig. 2. CIDOC CRM modelling of the MANN.2525 *cippus* and the VE.150 inscription

6 Conclusions and further work

This study has just scratched the surface of what is certainly a complex problem. Although the bases for creating an epigraphic extension of CIDOC CRM (CRMepi) have been laid, much of the work still remains to be done. If we consider that EpiDoc not only provides entities for the description of the text and its structural characteristics but also provides a series of tags for the identification of actor and place names inside it, for example, we must notice that the text itself may contain semantically relevant elements that need to be captured in some way. Actor appellations can for instance relate to the commissioners of a given monument or to the people to whom a certain epigraph was dedicated; furthermore, place appellations could also refer to places where the inscription was located in the past, or to which the text refers in various ways. These appellations may also evoke and prescribe the use of thesauri and gazetteers (like Pleiades, for example) to operate a further enrichment of the descriptions or for the creation of terminology tools starting from the very texts of the inscriptions. A special case, but one that is very frequent in epigraphy, is that of the so-called "talking objects", in which the inscription becomes a "declaration" made by the

object itself ("I am the cup of Aphrodite", "I was made by Ergotimos") that in this particular case acts as an "actor", as though it were a living entity. This type of data allows us to enrich our information network about the object and to expand our archaeological and historical knowledge.

Further developments will start from these bases to plan cycles of semantic enrichment of epigraphic information from textual data. For example, it will be possible to deduce and instantiate *E39 Actor* and *E53 Place* classes from the appellations found in the text, a process already attempted for discovering relevant semantic entities in ancient literary sources [15]. The use of XLink/XPointer technologies, also based on XML as EpiDoc, would make it possible to establish cross-references between semantic entities and specific portions of the epigraphic text.

7 Acknowledgements

The present work has been partially supported by the ARIADNE project, funded by the European Commission (grant 313193) under the FP7 INFRA-2012-1.1.3 call.

References

1. EpiDoc: Epigraphic Documents in TEI XML, http://sourceforge.net/p/epidoc/wiki/Home/
2. Guarducci, M.: Epigrafia greca. Caratteri e storia della disciplina. La scrittura greca dalle origini all'età imperiale. Istituto poligrafico e Zecca dello Stato, Roma (1967)
3. Panciera, S.: What Is an Inscription? Problems of Definition and Identity of an Historical Source. Zeitschrift für Papyrologie und Epigraphik 183, pp. 1–10 (2012)
4. Pulgram, E.: The typologies of writing systems, in Haas W. (ed.) Writing without letters, Mount Follick Series, vol. 4, Manchester University Press, pp. 1–28 (1976)
5. Pierce, C. S.: (1931-1958), Collected Papers of Charles Sanders Pierce, Burks A W (ed.), Harvard University Press, Cambridge, MA
6. Harris, R.: The Origin of Writing (1986), London, Duckworth
7. Krummrey, H., Panciera, S.: Criteri di edizione e segni diacritici, Tituli 2 (1980), pp. 205-215.
8. Doerr, M., Schaller, K., Theodoridou, M.: Integration of Complementary Archaeological Sources. In Nicolucci, F., Hermon, S. (eds.) Beyond the Artifact. Digital Interpretation of the Past. Proceedings of CAA 2004, pp. 64–69. Archaeolingua, Budapest (2010), http://proceedings.caaconference.org/files/2004/09_Doerr_et_al_CAA_2004.pdf
9. Lamé, M., et al.: Asking Text-Bearing Objects: Contribution of Epigraphical Theories to Digital Representation of Text, paper presented at SDH 2011, http://crdo.up.univ-aix.fr/SLDRdata/doc/show/copenhagen/SDH-2011/submissions/sdh2011_submission_48.pdf
10. EAGLE Project, http://www.eagle-network.eu/
11. ARIADNE Project, http://www.ariadne-infrastructure.eu/

12. CRMarchaeo, http://www.ics.forth.gr/index_main.php?l=e&searchtype=st_p&c=711
13. Vetter, E.: Handbuch der italischen dialekte, C. Winter, Heidelberg (1953)
14. Mommsen, Th.: Die unteritalischen Dialekte, G. Wigand, Leipzig (1850)
15. Andreussi, M., Felicetti, A.: The CIDOC CRM Encoding of the "Fontes ad Topographiam Veteris Urbis Romae Pertinentes" by Giuseppe Lugli. In: Posluschny, A., Lambers, K., Herzog, I. (eds.) Layers of Perception. Proceedings of CAA 2007. Dr. Rudolf Habelt GmbH, Bonn (2008), http://proceedings.caaconference.org/files/2007/66_Andreussi_Felicetti_CAA2007.pdf

Temporal Primitives, an Alternative to Allen Operators

Manos Papadakis and Martin Doerr

Foundation for Research and Technology - Hellas (FORTH)
Institute of Computer Science
N. Plastira 100, Vassilika Vouton, GR-700 13
Heraklion, Crete, Greece
{mpapad,martin}@ics.forth.gr
http://www.ics.forth.gr

Abstract. Allen Interval Algebra introduces a set of operators, which describe any possible temporal association between two valid time intervals. The requirement of Allen operators for complete temporal knowledge goes against the monotonic knowledge generation sequence, which is witnessed in observation driven fields like stratigraphy. In such cases, incomplete temporal information yields a disjunctive set of Allen operators, which affects RDF reasoning since it leads to expensive queries containing unions. To address this deficiency, we introduce a set of basic temporal primitives, which comprise the minimum possible and yet sufficient temporal knowledge, between associated time intervals. This flexible representation can describe any Allen operator as well as scenarios with further temporal generalization using logical conjunctions. Furthermore, an extension to the basic set of primitives is proposed, introducing primitives of improper inequality, which describe scenarios with increased imprecision that reflect disjunctive temporal topologies. Finally, the proposed temporal primitives are employed in an extension of CIDOC CRM.

Keywords: Temporal Primitives, Temporal Topology, Allen Operators, Incompleteness, Imprecision, CIDOC CRM, Knowledge Representation

1 Introduction

The substance of the past is considered as a set of phenomena [4] that manifested before a given point in time. For instance, a past era such as the *Minoan Period* comprises a set of cultural phenomena related to the Minoan civilization. The CIDOC conceptual reference model (CRM) [1] refers to the constituents of the past as temporal entities that cover a finite and continuous time frame over the timeline. Apart from the temporal facet, phenomena are also framed by a context that reveals their modeling purpose. Although temporal entities are regarded as interdependent wholes, the study of the past includes not only the sufficient description of the phenomena but also the relations among them, either semantic or temporal.

Since the past is not directly observable, knowledge about past phenomena is gained through the observation process of the available evidence, which justifies their existence [5]. Either the time-extent or the semantics that describe a set of phenomena can imply a possible temporal topology that holds among them (see also Chapter 3.3 in [9]). The prevailing method for representing temporal knowledge is Allen Interval Algebra [2]. This theory proposes a model that portrays the notion of time interval, along with a set of temporal relations, called Allen operators, which describe any possible temporal topology.

However, observation-driven fields such as stratigraphy, often extract vague and sparse information about the modeled temporal entities. Consequently, the use of Allen operators is hindered because they are bounded to the requirement of complete temporal knowledge. In addition, there are several cases in which semantic associations between coherent phenomena can reveal only a fraction of their possible temporal relation. Therefore, the representative Allen operators that describe the concluded association form a set of possible alternative scenarios of temporal relations, which blurs the total image.

The rest of this document is organized as follows. First, we provide an extensive description of the aforementioned concerns and a deeper analysis of the resulting issues. In Section 3 we address these issues by proposing a set of temporal relations able to describe any scenario that is constituted by incomplete and imprecise temporal knowledge. Finally, we analyze the expressiveness of the proposed relations, followed by some concluding remarks.

2 Background and Motivation

The main information components that frame a *temporal entity* include the context and the time extent. The semantics that frame a temporal entity i.e. interactions of things, people and places, determine its context, whereas the temporal projection confines the modeled phenomenon's extent over time. Although the distant nature of these information components, they are interrelated, resulting into relative inference. More specifically, relationships that hold between interactions within the content of the associated entities i.e. causal relation (cause and effect association) can reveal temporal dependency.

In order to illustrate the aforementioned concept, we focus on the notion of activity. According to CIDOC CRM [1], an activity represents a special case of a temporal entity, in which the included phenomena are considered as the outcome of intentional actions. Based on the context that introduces the semantics of the activity's instances, it is possible for a semantic association to hold, which in turn determines their possible temporal relations.

Semantic association between instances of activities is frequently referenced in literature. A common incident of logical connection is the case of influential correlation between activities. This type of phenomena has been encountered multiple times in fields related to the study of the past. The most common scenario refers to the case in which an activity instance determines the context of another individual instance. As a result, the coherent entities subjected to an

intentional continuation in time, implying that the latter instance is a consequence of the former. For instance, consider the recitation and the stenography of Homer's epic poems [7] from the spoken words of a singer to the manuscripts of a stenographer, respectively. The recording is regarded as a continuation in time of the narration activity in order to achieve the poems' preservation from the oral tradition to the written form.

The continuation phenomenon reveals the following reasoning chain: the influential association implies a continuation in time, which in turn entails a relevant temporal order between the activities. More specifically, considering the temporal constraints that enable a continuation phenomenon, it is intuitively proven that an activity instance cannot continue another instance that takes place in the future. With respect to the aforementioned statement, it is obvious that the recording activity cannot continue the narration activity, if the latter instance occurs after the former.

The temporal topology between related entities as well as the temporal constraints that describe a continuation phenomenon are instances of temporal information. Allen Interval Algebra [2] is an established means of representing such knowledge. According to Allens theory, a time interval is considered as an ordered set of points that represents a time frame on the timeline. Each time interval is considered as a continuous spectrum and is formalized by a pair of endpoints that indicate the starting and ending time point of the corresponding frame. It is worth noting a time interval is identified as valid if it conforms to a basic temporal constraint, which states that the interval cannot have zero duration i.e. its starting point must always be before the ending point.

Temporal constraints are considered as rules that describe a temporal relation; particularly, they associate the endpoints of the related intervals. Allen's theory introduces a set of temporal operators, known as Allen operators, that represent the possible relations between time intervals. The operators are formalized using a set of temporal constraints that associate all possible pairs of endpoints of the related intervals. For instance, operator *meets* represents the temporal relation of a meeting in time. The rules that describe this operator express that the end of a time frame signifies the start of the other. A detailed analysis of the Allen operators and the corresponding required endpoint constraints are presented in [10].

Allen Interval Theory [2] can be used to formalize the temporal topology that constitutes the continuation phenomenon. Let A and B denote the time intervals which represent the time extent of the "narration" and "recording" activities. Each interval is described by a set of temporal endpoints; As, Ae and Bs, Be depict the extreme points of interval A and B, respectively. Note that s stands for the starting point of an interval whereas e refers to the ending point. With respect to the latter notation, the continuation phenomenon is formalized as $As < Be$ which states that the start of the "narration" activity must occur before, in time, the end of the "recording" activity. For the sake of simplicity, from this point onward, any reference to time intervals A and B or their endpoints will also refer to the corresponding activity instances, unless explicitly stated

otherwise. As a result, a reference to As states the starting point of the time interval that represents the time extent of the "narration" activity and hence, the starting time of the activity itself.

The concluded endpoint constraint depicts the minimum temporal information that implies a continuation phenomenon. However, the corresponding temporal topology that may hold between the associated activities is efficiently described using Allen temporal operators, as it was mentioned above. Particularly, the endpoint constraint $As < Be$ reflects a set of probabilistic equivalent temporal relations that associates activity A and B as follows: A (is) {*before, meets, overlaps, overlapped-by, starts, started-by, during, includes, finishes, finished-by, equals*} B, in terms of Allen operators.

Every temporal relation that holds between the associated activities is applied in a disjunctive manner. Therefore, the resulting operators are connected with the logical operator *OR*. This operator emerges from the difference between the requirement of complete temporal knowledge that characterizes the Allen operators and the temporal incompleteness that is intertwined with the study of the past. Although disjunctive temporal information does not affect the expressiveness of Allen operators, it goes against the monotonic knowledge generation sequence. This contradiction raises both theoretical and practical issues.

On the one hand, an attempt to theoretically approximate the temporal topology of a scenario with notable incomplete knowledge, such as continuation in time, results to a set of twelve possible Allen operators. Although the exclusion of the single operator *after* is undoubtedly considered as knowledge and supports deductive reasoning, the remaining options still provide a blurry image of twelve possible interpretations. As far as technical aspects are concerned, the aforementioned possible scenarios have a significant effect on RDF reasoning. More specifically, the concluded set of Allen operators leads to expensive queries that contain unions of selection clauses, each of which expresses an alternative temporal association.

Incomplete temporal knowledge is widely witnessed in observation-driven fields, where completeness can only be achieved through consecutive information disclosure. For instance, in the field of stratigraphy [6] the logical association between layers of soil reveals a sequence of phases that manifested through time upon a specific geographic area. Temporal incompleteness among the starting and ending endpoints of the different layers is a scenario frequently described as a set of possible associations. The need for a more flexible representation of temporal topology emerges. In the following section, we propose a new temporal algebra, as an alternative to Allen Interval Theory, that combines existing knowledge in a conjunctive way, supporting monotonic knowledge gain and offers a basis for the efficient description of scenarios with temporal incompleteness.

3 Temporal Primitives

Considering that temporal imprecision is an inevitable characteristic that accompanies the description of past phenomena, our approach of representing the

temporal topology relies on the model of fuzzy intervals that was introduced in our previous work [8]. According to this model, temporal information of an interval is depicted as an aggregation of two sets of time point: the boundary set that represents a fuzzy layer within which the true endpoints are confined, and the interior set, which comprises the body of the interval. Consequently, the starting and ending endpoint of a fuzzy interval is represented by the lower and upper boundary set, respectively [8].

Based on the above fuzzy model, a meeting in time is no longer perceived as an endpoint equality, as introduced by Allen, but as an overlapped boundary zone. In addition, the ordering relations, which are depicted as endpoint inequalities in Allen's model are interpreted as ordering between ordered time point sets. For instance, the basic constraint that the start of an interval is *before* its end is expressed by requiring that every time point of the lower boundary set is before (in time) every time point of the upper one. From this point onward any reference to a time interval corresponds to its fuzzy representation, unless stated otherwise. Particularly, any reference to the endpoints of an interval implies the corresponding lower or upper boundary set, while endpoint equality and ordering are interpreted as described above.

In order to address the issue of temporal incompleteness, as analyzed in Section 2, we propose a set of primary temporal associations, applicable to fuzzy time intervals. In the remainder of this section we define the notion of temporal primitives and proceed to the introduction of seven basic relations, which are then extended to include four generalized primitives. Then we provide a visual representation of each temporal primitive using fuzzy intervals.

3.1 Basic Primitives of Equality and Proper Inequality

We define the notion of **temporal primitives** as a set of relations, which comprise the minimum possible and yet sufficient temporal knowledge, which describes the temporal topology that may hold between associated time intervals. Each primitive refers to the simplest, plausible relative association between pairs of endpoints in terms of temporal constraints, similar to those that form the Allen's operators. Note that the endpoint equality and the temporal ordering are considered as fuzzy interpretations, as it is explained in Section 3.

The core of each temporal primitive is an endpoint constraint, which is composed of two operands and a comparative operator. The operands are the endpoints of the intervals, while the operator is either "less than" or "equals to", representing the relations *before* i.e. temporal ordering, and (endpoint) *equality*, respectively. Although "greater than" is also a comparative operator, it is skipped, since its semantics correspond to an inversed "less than" relation.

According to the representative endpoint constraint, a temporal primitive describe either a generalized state of temporal topology i.e. a disjunction of possible Allen operators, or a specific temporal relation. Conjunctions of temporal primitives form temporal associations that reflect shorter sets of Allen operators.

Let A and B be two time intervals with endpoints (As, Ae) and (Bs, Be) respectively. Using the absolute operators of "equality" (=) and "less than"

($<$) we form seven basic temporal primitives, as shown below, along with the representative endpoint constraint and the corresponding set of Allen operators.

- **A starts before the start of B**: the starting endpoint of interval A occurred before the start of B. The representative endpoint constraint ($As < Bs$) corresponds to the following set of Allen's operators: A (is) *before OR meets OR overlaps OR includes OR finished-by* B.
- **A starts before the end of B**: the starting endpoint of A occurred before the end of B. The representative endpoint constraint ($As < Be$) corresponds to the Allen's operator set: A (is) *before OR meets OR overlaps OR starts OR started-by OR includes OR during OR finishes OR finished-by OR overlapped-by OR equals* B.
- **A ends before the start of B**: the ending endpoint of A occurred before the start of B. The representative endpoint constraint ($Ae < Bs$) is expressed as A (is) *before* B.
- **A ends before the end of B**: the ending endpoint of A occurred before the end of B. The representative endpoint constraint ($Ae < Be$) is expressed as A (is) *before OR meets OR overlaps OR starts OR during* B.
- **A starts at the start of B**: the starting endpoint of A occurred at the start of B. The representative endpoint constraint ($As = Bs$) is expressed as A (is) *starts OR started-by OR equals* B.
- **A ends at the start of B**: the ending endpoint of A occurred at the start of B. The representative endpoint constraint ($Ae = Bs$) is expressed as A *meets* B.
- **A ends at the end of B**: the ending endpoint of A occurred at the end of B. The representative endpoint constraint ($Ae = Be$) is expressed as A (is) *finishes OR finished-by OR equals* B.

3.2 Generalized Primitives of Improper Inequality

The synthesis of the basic temporal primitives relies on the exhaustive combination of endpoint constraints that are formed using absolute operators, that is, "equality" or "less than". However, temporal imprecision is not only witnessed in the definition of a time interval (fuzziness when discovering the past) but also in the semantics itself. For instance, negative evidence between temporal entities may lead to the negation of an *after* relation, which in turn reveals an imprecise continuation in time expressed as a disjunction of absolute temporal constraints.

Since, the previous example cannot be expressed with a single or a conjunction of absolute operators, we need to introduce an additional *generalized* operator, "less than or equal" ($\leq$), which describes the temporal constraint of *before or equal* (in time) i.e. improper inequality (see also Chapter 3.4 in [9]). Using this operator, we propose four additional temporal primitives that represent disjunctive combinations of the basic primitives. Let A and B be two time intervals with endpoints (As, Ae) and (Bs, Be) respectively. Using the improper inequality operator ($\leq$), we propose the following generalized primitives.

- **A starts before or at the start of B**: the starting endpoint of interval A occurred before or at the start of B. The representative endpoint constraint ($As \leq Bs$) is expressed as A (is) *before OR meets OR overlaps OR starts OR started-by OR includes OR finished-by OR equals* B.
- **A starts before or at the end of B**: the starting endpoint of A occurred before or at the end of B. The representative endpoint constraint ($As \leq Be$) is expressed as A (is) *before OR meets OR met-by OR overlaps OR overlapped-by OR starts OR started-by OR includes OR during OR finishes OR finished-by OR equals* B.
- **A ends before or at the start of B**: the ending endpoint of A occurred before or at the start of B. The representative endpoint constraint ($Ae \leq Bs$) is expressed as A (is) *before OR meets* B.
- **A ends before or at the end of B**: the ending endpoint of A occurred before or at the end of B. The representative endpoint constraint ($Ae \leq Be$) is expressed as A (is) *before OR meets OR overlaps OR starts OR during OR finishes OR finished-by OR equals* B.

Figure 1 illustrates the temporal relations of intervals A and B, using the basic and generalized primitives. It is worth noting that, due to limited space, the figure excludes extreme cases; the interested reader can refer to [10].

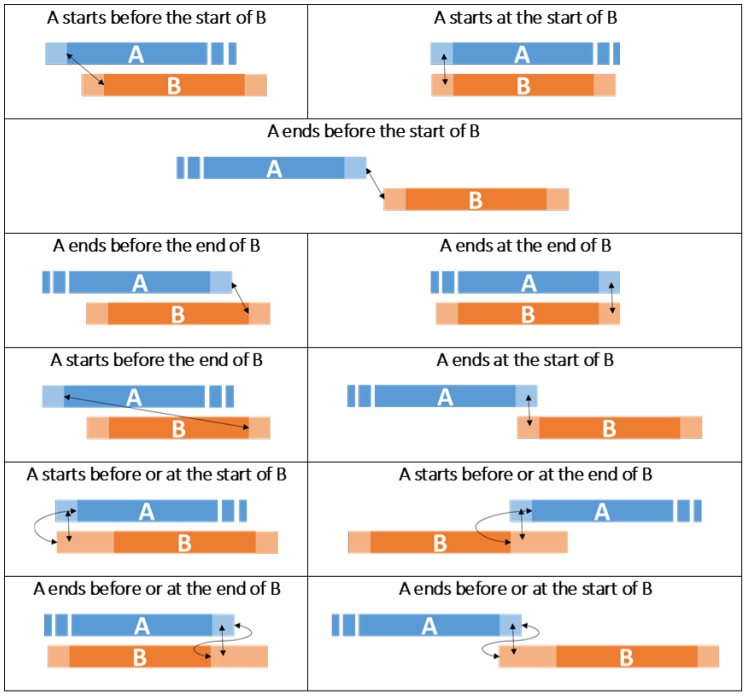


Fig. 1. Temporal Primitives

4 Expressive Power of Primitives

In this section, we analyze the expressiveness and flexibility of the temporal primitives. First we focus on the completeness and minimality that characterizes them. Then, we introduce a subsumption hierarchy graph that organizes primitives based on their expressiveness. Finally, we provide a complete representation of each Allen's operator exclusively using temporal primitives.

4.1 Completeness and Minimality

Since the proposed set of temporal primitives is an alternative to Allen's operators, it must be complete and yet minimal. Table 1 shows that every possible endpoint constraint can be expressed using exclusively primitives; the interested reader can refer to [10].

Table 1. Temporal Primitives Completeness and Minimality

Endpoint Constraint	**Temporal Primitive**
As $<$ Bs	A starts before the start of B
As $\geq$ Bs	B starts before or at the start of A
As $<$ Be	A starts before the end of B
As $\geq$ Be	B ends before or at the start of A
Ae $<$ Bs	A ends before the start of B
Ae $\geq$ Bs	B starts before or at the end of A
Ae $<$ Be	A ends before the end of B
Ae $\geq$ Be	B ends before or at the end of A
As $=$ Bs	A starts at the start of B
Ae $=$ Bs	A ends at the start of B
Ae $=$ Be	A ends at the end of B

4.2 Subsumption Hierarchy Graph

The expressive power of each primitive is subjected into a hierarchical structure, in which primitives with stronger interpretations subsume weaker ones. Figure 2 organizes the temporal primitives based on their expressiveness. Note that the dashed boxes refer to representative set of Allen's operators. The upper levels of the graph refer to generalized temporal topologies, while lower levels describe specific relations. This structure grants flexibility, allowing efficient reasoning in cases of information revision. For instance, given a certain set up of temporal knowledge it is concluded that two activities are associated with a "starts before the start of" relation. Following the graph it is straightforward to resolve which can be the possible temporal topology in the case of weakening or strengthening the endpoint constraints. On the contrary, Allen's operators are not subjected into a hierarchy since no subsumption relations exist among them.

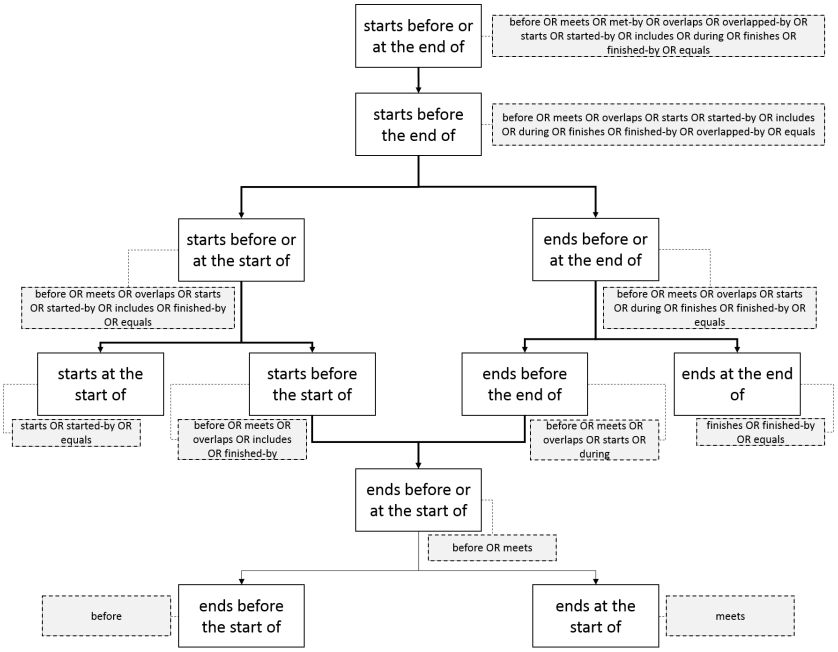


Fig. 2. Hierarchy Graph

4.3 Allen Alternative Representation

Each temporal primitive represents the simplest form of a temporal relation, which refers to an endpoint constraint that relates the intervals under consideration. Since Allen operators are built by combining several meaningful endpoint constraints, it is intuitively implied that a proper sequence of representative primitives can describe every Allen relation as well. Table 2 offers a corresponding conjunctive combination of temporal primitives for each Allen's operator.

Table 2. Allen operators expressed as Temporal Primitives

Allen operator	**Temporal Primitives**
before	A ends before the start of B
meets	A ends at the start of B
overlaps	A starts before the start of B *AND* B starts before the end of A *AND* A ends before the end of B
starts	A starts at the start of B *AND* A ends before the end of B
during	B starts before the start of A *AND* A ends before the end of B
finishes	B starts before the start of A *AND* A ends at the end of B
equals	A starts at the start of B *AND* A ends at the end of B

5 Conclusion

This paper proposes a set of temporal primitives as a flexible alternative to Allen's operators, which efficiently describe temporal topologies characterized by incomplete knowledge and imprecision. The proposed relations rely on the Fuzzy Interval Model [8] in order to express imprecise temporal knowledge that is witnessed in observation-driven fields. Each primitive encapsulates the expressiveness of a simple yet plausible endpoint constraint, similar to those that built the Allen Interval Algebra. The set of temporal primitives conforms to the principles of completeness and minimality, while their expressiveness allows for a hierarchical association among them.

This study resolves the problem of temporal knowledge representation using disjunctive Allen operators, as expressed in issue 195 of CIDOC SIG [3]. The proposed temporal primitives have been introduced as scope notes in the definition of CIDOC CRM, in order to represent properties of class *E2:Temporal Entity* that subsume the corresponding Allen operators. An extended analysis of this work can be found in [10].

References

1. Definition of the cidoc conceptual reference model version 5.0.4. Tech. rep., ICOM/CIDOC CRM Special Interest Group (11 2011)
2. Allen, J.F.: Maintaining knowledge about temporal intervals. Communication of ACM pp. 832–843 (1983)
3. CRM, C.: Crm. the cidoc conceptual reference model (iso/cd21127)http://http://www.cidoc-crm.org/ (2013)
4. Doerr, M., Kritsotaki, A., Stead, S.: Which period is it? A Methodology to Create Thesauri of Historical Periods. In: Proceedings of the Computer Applications and Quantitative Methods in Archaeology Conference (CAA2004). pp. 13–17. Prato, Italy (April 2004)
5. Doerr, M., Plexousakis, D., Kopaka, K., Bekiari, C.: Supporting chronological reasoning in archaeology. In: In Computer Applications and Quantitative Methods in Archaeology Conference, CAA2004. pp. 13–17 (2004)
6. Gradstein, F.M., Ogg, J.G., Smith, A.G.: Chronostratigraphy: linking time and rock. In: Gradstein, F.M., Ogg, J.G., Smith, A.G. (eds.) A Geologic Time Scale 2004, pp. 20–46. Cambridge University Press (2005)
7. Lord, A.B.: The Singer of Tales. Harvard University Press (2000)
8. Papadakis, M., Doerr, M., Plexousakis, D.: Fuzzy times on space-time volumes. In: eChallenges e-2014, 2014 Conference. pp. 1–11 (Oct 2014)
9. Papadakis, M.: Temporal Topology on Fuzzy Space Time Volumes. Master's thesis, Computer Science Department, University of Crete (2014)
10. Papadakis, M., Doerr, M.: Temporal primitives. Tech. rep., Foundation for Research and Technology (FORTH) (2015)