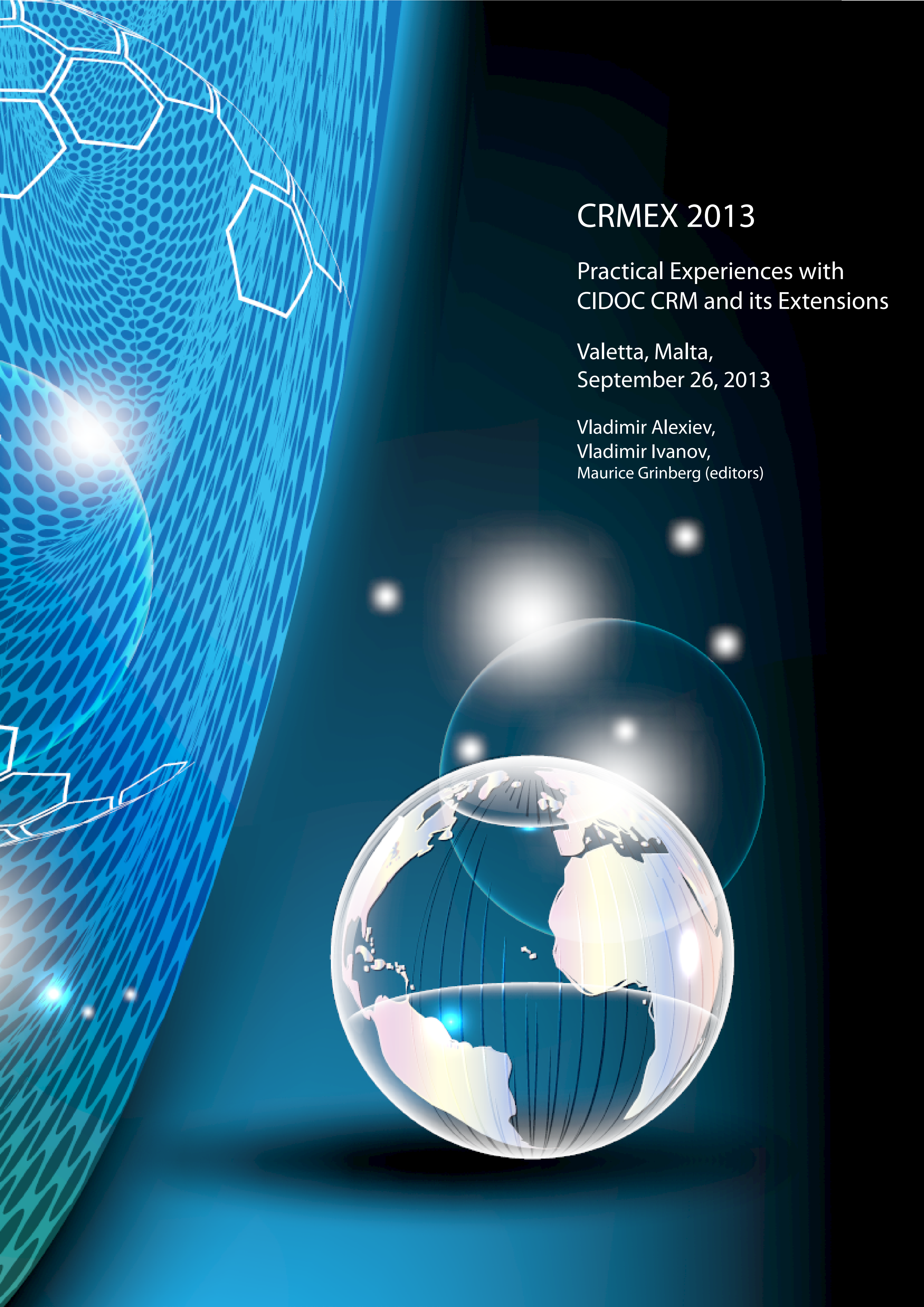


CRMEX 2013

Practical Experiences with CIDOC CRM and its Extensions

Valetta, Malta,
September 26, 2013

Vladimir Alexiev,
Vladimir Ivanov,
Maurice Grinberg (editors)

Contents

Colophon

Cite as follows:

Vladimir Alexiev, Vladimir Ivanov, Maurice Grinberg (editors),
Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), 26 September 2013, Valetta, Malta. Papers and presentations available online at http://www.ontotext.com/CRMEX and CEUR WS.

Published December 2013

Foreword

The workshop **Practical Experiences with CIDOC CRM and its Extensions (CRMEX)** ran on 26th September 2013 in Valetta, Malta, as part of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013).

The CIDOC CRM (international standard ISO 21127:2006) is a conceptual model and ontology with a fundamental role in many data integration efforts in the Digital Libraries and Cultural Heritage (CH) domain. It has spawned various CRM-compliant extensions, such as:

- Functional Requirements for Bibliographic Records (FRBRoo) for works and bibliographic data;
- CRM Digitization (CRMdig) for digitization and provenance;
- CRM for English Heritage (CRMEH) for archaeology;
- British Museum Ontology (BMO) for museum objects;
- Sharing Ancient Wisdoms (SAWS) for medieval gnomologia (collections of wise sayings);
- PRESSoo, a FRBRoo extension for serial publications.

A number of data models, while not CRM-compliant, have been influenced by the CRM, e.g. the Europeana Data Model. At the same time, some people claim that the examples of practical working systems using CRM are few and far between. There are various difficulties facing wider CRM adoption and interoperation, e.g.:

- Because CRM allows many different ways of representing the same situation, CRM adopters in various CH areas need mapping guidelines and best practices to increase the chance of interoperation;
- While RDF is the most viable CRM representation, there are various low-level RDF issues that are not standardized. Since RDF representation implies a certain implementation bias and still undergoes changes of good practice, CRM-SIG has been expecting good practices to emerge from people applying CRM in order to make recommendations.

The goal of this workshop is to describe and showcase systems using CRM at their core, exchange experience about the practical use of CRM, describe difficulties for the practical application of CRM, and share approaches for overcoming such difficulties.

The ultimate objective of this workshop is to encourage the wider practical adoption of CRM.

Topics

- Software systems and similar developments using CRM;
- CRM repositories that aggregate large amounts of CRM RDF data;
- CRM-compliant extension ontologies and domain specializations. Principles for extending CRM;
- Best practices for representing specific situations from specific CH domains in CRM;
- Best practices, guidelines and detailed mappings from various metadata formats and various CH domains to CRM;
- Joint use of CRM and other popular ontologies. Principles for selecting constructs from different ontologies;
- Querying, searching and faceted browsing of CRM repositories;
- Display, editing, annotation and cross-linking of CRM data;
- Reasoning with CRM data;
- Encountered mistakes in representing CRM data. CRM learning curve and didactic considerations;
- Shortcomings of CRM, recommendations for CRM evolution. Collaboration on CRM evolution, merging RDF standardization approaches, recommendations for collaborative approaches;
- Performance and volumetric information about CRM-based systems;
- Evaluations of CRM adoption, usability of CRM-based systems, usage of specific CRM constructs.

Organizing Committee

- Vladimir Alexiev, Ontotext, Bulgaria, Workshop Chair
- Vladimir Ivanov, Kazan Federal University, Russia, Review Chair
- Maurice Grinberg, New Bulgarian University, Publication Chair
- Christian-Emil Ore, University of Oslo, Norway, Authors Liaison
- Guenther Goerz, University of Erlangen-Nuremberg, Germany, Publicity Chair

Program Committee

- Ceri Binding, Hypermedia Research Unit, Hypermedia Research Unit; Faculty of Computing, Engineering and Science; University of South Wales
- Christian-Emil Ore, Unit for Digital Documentation, University of Oslo, Norway
- Costis Dallas, Associate Professor, Director of Museum Studies, University of Toronto
- Eero Hyvönen, Professor and Research Director, Semantic Computing Research Group, Department of Media Technology, Aalto University and University of Helsinki
- Franco Niccolucci, Director, VAST-LAB, PIN, University of Florence, Prato, Italy (former: Professor at the Faculty of Architecture)
- Kai Eckert, Postdoctoral Researcher, Research Group Data and Web Science, University of Mannheim
- Martin Doerr, Research Director, Center for Cultural Informatics, Information Systems Laboratory, Institute of Computer Science, FORTH, Greece
- Michele Pasin, Information architect, Nature Publishing Group (formerly research associate, Department of Digital Humanities, King's College)
- Øyvind Eide, Senior Analyst, Unit for Digital Documentation, University of Oslo, Norway
- Patrick Le Boeuf, Bibliothèque nationale de France
- Rainer Simon, senior researcher, Digital Memory Engineering research group, Austrian Institute of Technology
- Stefan Gradmann, Professor at Faculty of Arts, Director of the University Library, Katholieke Universiteit Leuven
- Trond Aalberg, Associate Professor, Data and Information Management group, Norwegian University of Science and Technology
- Vladimir Alexiev, Lead, Data and Ontology Management group, Ontotext Corp, Bulgaria
- Vladimir Ivanov, Senior research assistant, Computational Linguistics Laboratory, Kazan Federal University, Russia. Cultural Heritage Digitization Center of Tatarstan

A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts

Martin Scholz

Universität Erlangen-Nürnberg,
Erlangen, Germany
martin.scholz@fau.de

Abstract. The detection of event mentions in free text is a key to a deeper automatic understanding of the text's contents. In this paper we present ongoing work on mechanisms to detect events in German texts in the domain of cultural heritage documentation. A central role plays a hand-crafted mapping of CIDOC CRM[1] events to GermaNet synsets to ease the process of creating a lexicon for automatic event detection. We discuss two approaches and insights gained from the mapping process and correct modelling of event mentions.

1 Introduction

In cultural heritage, free text is an important source of information and a popular form of documentation. For the latter, free text is often combined with structured metadata records. While the records provide basic, standardized metadata, the texts contain more detailed descriptions or additional information. Structured metadata can be accessed and processed quite well by machines For the contents of free text, however, this does not hold. Although there exist various methods for automatic information extraction, currently none can reach the high quality of expert-proven data necessary for academic research. Their efficacy varies heavily with text properties such as language, genre, etc; and this most likely will not change in near future. It is therefore desirable to semantically enrich texts with human revised annotations in order to extract its contents in a machine-processable way with quality sufficient for scolarly research.

One approach is to assist human annotators with automatic text analysis methods, providing for annotation proposals. Such an approach is implemented in the WissKI system, as described in Sect. 3.3.

Basically, such detection algorithms rely on one of two types of data resources for computing their heuristics: Either on a large-scale annotated corpus or on a (hand-made) lexicon. For common named entity classes like persons, places, organisations and times there are hand-annotated corpora and ready-to-use automatic annotation tools available, although languages other than English are supported much more rarely [14], [4], [3]. Events[2] are covered less frequently. The

[1] In this paper we always refer to version 5.0.4 of the CIDOC CRM [2].

[2] Note that the notion of what the term "event" means varies in information retrieval. E.g. some literature focuses rather on (historical) periods like "industrialisation". In this paper we align our understanding of the term with the class *E5 Event* in the CRM.

Timebank corpus[3] [8], an English corpus annotated with TimeML[4] [1] mark-up language, also contains annotations of events and there is some literature about event detection [9]; again, mostly for English. For our target language, German, we are currently not aware of any freely available corpus with event annotations or tools for automatic event detection.

In this paper we describe a mapping of CIDOC CRM event classes to GermaNet, a wordnet for the German language. From the mapping and GermaNet, a word list can be compiled that is the basis for an event detection algorithm. As we are not aware of available German corpora tagged with CIDOC CRM event classes, we also built a small manually annotated corpus of text from museum documentation, which we use for development and evaluation.

The rest of the paper is structured as follows: First, the lexical resource for our mapping, GermaNet, is briefly described. In the following section we present a simple and a more elaborate mapping strategy and shortly discuss their strengths and weaknesses. Then, the detection algorithm is described and evaluated against a small hand-crafted corpus. Further, we describe its application in the WissKI system. In Sect. 4 observations and challenges for future work are discussed. Finally, we conclude with Sect. 5.

1.1 German Wordnet

GermaNet[5] [6] is a German wordnet. Its structure is based on the Princeton WordNet[6] for English. Unlike Princeton WordNet, GermaNet is not open data, but only free for academic research. The work described here is based on GermaNet version 7.0.

Key concept of the family of wordnets is the so-called synset, a set of words[7] that are synonyms in a certain textual context. A synset is thus an equivalence class, i.e. the words of a synset can be used interchangeably in that context.

A word can participate in several synsets, reflecting large or small shifts in its meaning. The meanings of a word are numbered, so that a specific meaning — a so-called word sense — can be identified by the word and an integer. Likewise, a synset can be identified by the word sense of one of the words it contains. GermaNet distinguishes three parts of speech or word categories: noun, verb, and adjective.

Synsets are linked to each other by certain semantic relationships, like antonymy or meronymy. The predominant one is hypernymy. Synsets are usually arranged hierarchically according to the hypernym-hyponym relation, forming a thesaurus. A synset may have multiple hypernyms.

A synset can be regarded as resembling the common meaning of a set of words. Thus, a synset can be seen as the lexical equivalent to a concept in an ontology while the hypernymic relation corresponds to the subclass relation. In fact, there have been some proposals to model wordnets as ontologies [7].

[3] http://www.timeml.org/site/timebank/timebank.html

[4] TimeML is a vocabulary for annotating temporal expressions in text. See http://www.timeml.org

[5] http://www.sfs.uni-tuebingen.de/lsd

[6] http://wordnet.princeton.edu

[7] Strictly speaking, a synset contains one or more so-called lexical units. A lexical unit contains the uninflected word form with possible orthographic variants. To keep it simple, we will not distinguish "word" from lexical unit.

2 The Mapping Mechanism

The idea of using GermaNet for event detection is that the structure of GermaNet can be exploited to generate large lists of words identifying CRM events by mapping an event class to a handful of synsets, rather than generating a list of words by hand. We assume that if the words of a synset can be used to denote a CRM event class, then its hyponyms are likely to also support this class. The more hyponyms the synset has, the more words can be selected with relatively small effort.

In this section we present two approaches for such a mapping[8] for CRM *E5 Event* and its subclasses, with two exceptions: *E13 Attribute Assignment* and its subclasses were not taken into account, as a first examination of the corpus data indicated that instances of this class are preferably expressed grammatically differently from other event classes. This may be due to the generic, metalevel-like nature of *E13 Attribute Assignment*. *E87 Curation Activity* was excluded primarily as it was out of scope of our research, but also because we were unsure about its extent and what words support it.

2.1 A Simple Approach

We first implemented a naive mapping approach. For each event class a small set of synsets was determined with two conditions:

1. the synset supports the concept
2. all hypernymic synsets do not support the concept

A synset supports a concept if one of its word senses refers to the class. Note that it is not required that each word sense of a word must refer to the class. Figurative use of words was not taken into account.

The second condition brings about that only the topmost synsets (in the sense of hypernymy) relatable to that event class are chosen, leading to a minimal set of synsets.

With appropriate tools for exploring the synset graph like GermaNet Explorer[9] such a mapping was built quite rapidly.

The mapping rules are expressed in XML:

```
<class name="ecrm:E67_Birth">
  <synset pos="v" word="gebären" sense="1" />
  <synset pos="n" word="Geburt" sense="1" />
  <synset pos="n" word="Geburt" sense="2" />
  <synset pos="n" word="Geburt" sense="3" />
</class>
```

Fig. 1. Declaration of synsets mapping to the *E67 Birth* event

A conversion programme was developed that compiles the synsets to a list of words: First, all hyponymic synsets are fetched from GermaNet. Then, the words contained in the synsets are extracted and printed with their word category. Duplicates are omitted. The result is again an XML mapping of event classes to words as shown in Fig. 2.

[8] The second mapping approach is available as an XML file for downloaded from http://wiss-ki.eu/node/167.

[9] http://www.sfs.uni-tuebingen.de/lsd/tools.shtml#GermaNet-Explorer

```
<class name="ecrm:E67_Birth">
    <word lemma="gebären" pos="v"/>
    <word lemma="entbinden" pos="v"/>
    <word lemma="niederkommen" pos="v"/>
    <word lemma="werfen" pos="v"/>
    <word lemma="laichen" pos="v"/>
    ...
    <word lemma="Geburt" pos="n"/>
    <word lemma="Drillingsgeburt" pos="n"/>
    <word lemma="Niederkunft" pos="n"/>
    <word lemma="Entbindung" pos="n"/>
    <word lemma="Totgeburt" pos="n"/>
    ...
</class>
```

Fig. 2. Excerpt from the compiled word list for *E67 Birth*

2.2 Problems of the First Approach

This simple approach shows two shortcomings:
The first problem arises from the polysemy of words. A word with different meanings — and thus contained in different synsets — is less likely to actually denote a specific event class than a word with only one meaning. Also, one meaning might be more frequent than another.
The predominant problem with this first approach, however, is that the scope of a CIDOC CRM event and the meaning of GermaNet word senses and synsets virtually never match exactly, but rather overlap. So, although a synset may support a CRM event class, the words of an hyponymic synset, however, may in no case support the event. This is illustrated by two prominent cases:
In CRM, the *E67 Birth* event only holds for humans. The birth of other living beings like animals is modelled with *E63 Beginning of Existence*. The top synset "gebären" in Fig. 1 supports the notion of a human birth and its words are the most commonly used in German for such an event. But they also may denote an animal birth. Consequently, some lower synsets introduce words that cannot be applied (unless as a colloquial or pejorative term) to human births, like "werfen" (mostly used for mammals with a bunch of offspring) or "laichen" ("spawn")[10] as shown in Fig. 2.
Another special case arises from the CRM clearly dividing things into material (*E19 Physical Thing*) and immaterial (*E28 Conceptual Object* and *E90 Symbolic Object*). This also affects the CRM event classes, as there are different classes for both branches: e.g. *E12 Production*/*E11 Modification* vs. *E65 Creation*. The German language and thus GermaNet, however, do not reflect this division. As a result, it is hardly impossible to find sufficiently broad synsets for which all words and hyponyms support the event. Only synsets with specialized meaning and with no or very little hyponyms fulfill this criterion. Synsets with frequently used words like "erschaffen", "erzeugen", "produzieren" (create, produce) all contain a wild mixture of hyponymic synsets applicable to events affecting either material things or immaterial things or both.

[10] In some cases GermaNet seems to be inconsistent: While "werfen" and "laichen" are grouped as birth, bird reproduction words like "legen" (lay an egg) or "schlüpfen" (hatch) are not.

An option would be to change the policy described in the previous section and only select synsets with words which always imply the event class. However, this leads to significantly less synsets and often excludes the most commonly used words, like “gebären” from *E67 Birth*.

2.3 A more fine-grained mapping

To overcome the shortcomings of the first approach, the mapping was extended so that hyponymic synsets can be excluded from the compilation process. For the XML notation, two modes were defined:

1. The element `<exclude_synset>` references a single synset that will be excluded. Its descendants are also excluded unless they can be reached via another branch or by another selected synset.
2. The boolean attribute `descend` for the `<synset>` element controls whether hyponyms should generally be included or excluded for this very synset. If set to `false`, all hyponyms of a synset are excluded by default.

The latter is primarily for convenience. However, it can also be regarded to lower the degree of semantic overlap of the synset and the CRM class: If set to `true`, the overlap is deemed to be rather high, as hyponyms are included by default. Analoguously, when `false`, the overlap is rather low.

Sometimes, synsets should be included that were implicitly excluded by one of the two methods. In such a case, the synset is explicitly selected, i.e. added to the synset list just like a top synset. Fig. 3 shows two examples: The Birth event now excludes all verbs denoting animal reproduction. The *E66 Formation* event is mapped to the synset “Heirat” (wedding) which mainly contains other activities as hyponyms like wedding anniversaries that don’t support *E66 Formation*. Therefore, they are excluded by default. The hyponym “Liebesheirat” (“marriage for love”), however, is explicitly included.

```
<class name="ecrm:E67_Birth">
  <synset pos="v" word="gebären" sense="1">
    <exclude_synset word="werfen" sense="5" />
  </synset>
  ...
</class>

<class name="ecrm:E66_Formation">
  ...
  <synset pos="n" word="Heirat" sense="1" descend="false" />
  <synset pos="n" word="Liebesheirat" sense="1" />
  ...
</class>
```

Fig. 3. Synsets can be explicitly excluded from the mapping

Although the events affecting material or immaterial things can be mapped quite accurately, the mapping is still not optimal as a lot of excludes have to be defined: The *E11 Modification* event maps to five topmost synsets, but with about 200 exclude statements. In such cases, the mapping process becomes quite time-costly and error-prone as the whole subtree must be scanned for synsets to exclude.

The conversion tool was adapted accordingly. Furthermore, each word will be given a confidence value between 0 and 1 that resembles the confidence that the intended meaning or word sense of the word in the given context is one of the word senses denoting the event. It is computed as follows:

$$confidence(w) = \frac{s_{w,e}}{s_w}$$

$s_{w,e}$ is the number of word senses of word w contained in the mapping for event e and s_w is the total number of word senses for word w.

The confidence can be used by a parser to rank event findings. However, this value only very roughly approximates the actual frequency of word senses in human language or a corpus.[11]

3 Event Detection

The compiled word lists are used for list-based event detection in the cultural heritage domain. The texts are tokenized, lemmatized and tagged with parts of speech (POS) using the Stuttgart TreeTagger [11]. A small script resolves separable verb particles, i.e. adds a particle to the corresponding verb lemma. [12]

In order for a token to be annotated as denoting an event, its lemma must occur in the corresponding word list and the POS tag must match the word category. Tokens may be annotated with multiple event classes. However, only the most specialized classes are kept, i.e. if a token is annotated with *E9 Move* and *E7 Activity*, the latter one is discarded as it is implicit in the former one.

At the moment, the algorithm does not perform word sense disambiguation. Words are annotated with possible events for each word sense. However, event annotations can be ranked according to the confidence value mentioned above.

3.1 Light Verbs

In German, light verb constructions are frequent, especially in scientific writing. Light verb constructions consist of a verb and a noun phrase, usually a nominalized verb, sometimes also including a preposition. Within this construct the noun carries the overall meaning, while the verb is reduced to only add a certain aspect[13] like causation. Typical examples include “erfolgen” or “stattfinden” (“take place”) together with an event-baring noun and rather fixed or lexicalized collocations like “zum Einsturz bringen” (“cause to collapse”).

A lot of light verbs can also occur on their own with a distinguished meaning (e.g. “bringen" then meaning “to bring”) and as such may also denote an event.

[11] This could be done in a further step, though, by computing the word sense frequencies from a corpus annotated with word senses, like the WebCAGe corpus (http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/webcage.html).

[12] In German language, a separable verb particle is a part of a verb that may occur separated from the verb stem in a proposition. The particles usually change the meaning of the verb, leading to totally different event classes.

[13] In German linguistics the common term is Aktionsart. A light verb usually shifts the focus to a certain aspect of the event, like beginning, end, result or cause.

In contrast, a light verb does not denote an event. A parser ignorant to light verb constructions will therefore produce much more false positives.

We also included a lexicon-based postprocessor to detect light verb constructions. Our parser uses a small hand-crafted lexicon and a dependency parser[14] in order to find such constructions. For a match, the verb is stripped off any event annotations. Event annotations for the noun part are augmented with aspect information provided by the light verb. We expect the aspect information to be a valuable hint in the role labeling phase that we plan to implement and for the right event modeling (see Sect. 4.2).

3.2 Evaluation on a Small Annotated Corpus

The coverage of the mapping was tested on a small corpus of short texts about museum objects.[15] The texts were annotated with event mentions manually.

Currently, the corpus contains 50 annotated texts with over 3000 tokens and 500 annotations.

For evaluation, a found annotation would be considered relevant if the corpus contained an annotation with same event class and that had at least 50% overlap. Conversely, a relevant annotation would be marked as missed, if the parser's output would not contain an annotation that suffices these conditions.

We achieve a precision of 59% and recall of 72%.

3.3 Use in the WissKI System

Our event detection system is developed as a part of the WissKI[16] virtual research environment[17]. WissKI is web-based, extending the popular content management system Drupal. It consistently relies on semantic web technology. Data is stored according to the CIDOC CRM in its OWL-DL implementation Erlangen CRM[18].

In WissKI, one form of data acquisition consists of semantically annotating free text in a WYSIWYG editor [12], [5]. From the enriched text, RDF triples can then be generated automatically. Annotations include entities like persons, objects, places, calendar dates, and events, and relations between these entities. The annotation process is designed to be semi-automatically:[19] WissKI provides the user with multiple annotation proposals. The user may always edit machine-produced annotations. Thus it is more important for the system to compute a (ranked) list of possible annotations than a single best solution. From this follows immediately that a higher recall is more favourable than high precision.

[14] We use the dependency parser ParZu [13] from the University of Zürich http://kitt.cl.uzh.ch/kitt/parzu/.

[15] The texts describe European works of art and are part of the online presentation of the exhibition about Renaissance, Baroque, and the Age of Enlightenment by the Germanic National Museum, Nuremberg, Germany.

[16] The WissKI project was funded by the German Research Foundation (DFG) from 2009-2012. Since then the WissKI software has been further developed.

[17] http://wiss-ki.eu

[18] http://erlangen-crm.org

[19] We don't expect natural language processing techniques to become accurate enough to obtain high-quality annotations in the near future. Therefore, machine-generated annotations must be approved by human experts to garantee annotation quality that meets academic standards.

4 Further challenges

The work on CRM event mapping and detection has raised some issues that we want to address in the future.

4.1 Mapping to English Wordnet

For English, there are much more sources of annotated data, but also linguistic resources and tools for event detection than for German. Consequently, a similar mapping for the English Princeton WordNet could reveal interesting insights for event detection, also for German. The Interlingual Index[20], an outcome of the EuroWordNet project, serves to build mappings between wordnets of various languages by introducing an intermediate layer. The mapping between GermaNet and Princeton Wordnet is kept up-to-date by the makers of GermaNet.[21] It remains to be seen if it could serve as a starting point or it is better to start from scratch.

4.2 When is an event a CRM event and of what kind?

The detection of events is just a first step towards an accurate modelling of events according to the CIDOC CRM. In fact, an event annotation can be modelled quite differently in CRM, depending on the context:
The CRM only models events as *E5 Event* if they actually took place. Hypothetical events, instead, should be modelled as conceptual objects like *E55 Type* or *E29 Design or Procedure*.
Further, a word literally denoting a certain event class may be actually modelled as a superclass of the event. For example, this is the case for events expressed with words that usually denote specializations of *E7 Activity* like *E12 Production* or *E8 Acquisition*, but that have been interrupted and produced no result. An example from the corpus is
"[. . .] Dentatus weist die Geschenke [. . .] zurück."
"[. . .] Dentatus rejects the presents [. . .]"
where the implied transfer of ownership (to give a present) could not be completed, and thus is just an *E7 Activity*. Nonetheless, it is of importance to also model the intended action. Likewise, events normally supporting (sub)classes of *E63 Beginning of Existence* or *E64 End of Existence* may fall back to *E5 Event*.
It is also important to detect how many event instances a word evokes. Based on the data in the corpus, we differentiate three cases depending on the number of individual events that are referred to:

individual: the word refers to only one single event instance. In most cases this event can be modelled as CRM event, unless it is hypothetical.

collection: the word refers to multiple but distinguished event instances of the same class. As in the case of an individual the events can be modelled as CRM events.

class: the word refers to a class of events rather than to event instances. Often, processes are described and so appropriate CRM classes would be *E29 Design or Procedure* or similar — as with hypothetical events.

[20] http://www.illc.uva.nl/EuroWordNet/

[21] http://www.sfs.uni-tuebingen.de/lsd/ili.shtml

The border between collection and class can be blurred and hard to identify. A collection of events is usually linked to a description of a well-defined collection of items or group of people. A class usually co-occurs with terms denoting classes of items. Thus, the correct modelling of events is highly dependent of the entities in context.

For the right modelling grammatical numerus is an important clue. The singular invokes the individual case while the plural invokes the collection or class case. Also, key words like "solche" ("such"), "diese" ("these") and other determiners can help to distinguish a class from a collection.[22]

TimeML also addresses this issue by distinguishing between event tokens and event instances, for a collection or individual. Classes (called "generics"), however, are not treated by TimeML [1], [10, pp. 1–8, 32–35].

4.3 Implicit Events

As seen in the sentence "Dentatus rejects the presents" an event mention can be co-triggered by a word primarily referring to an object or person; in this case the word "presents", denoting the material things in first place, but also the mode of handing over. Other frequent words include "Maler" (painter), "Gemälde" (painting) and family relations like "Tochter" (daughter) or "Vater" (father), including a *E12 Production*, *E65 Creation* or *E67 Birth* event, respectively.

It is hard to draw a line if event classes should be co-triggered with a certain word and if so, which ones. While the aforementioned "Gemälde" clearly triggers an *E12 Production*, it is not so clear for "Kunstwerk" (*work* of art) and "Objekt" (object) would not — although "Gemälde" and "Objekt" are both hyponyms of "Kunstwerk".

We have no clear guidelines yet. Our current practice is that a word denotes an event if it was somehow morphologically derived from a word denoting that event.

Nevertheless, such information can help in finding the right relation for constructions like
"Albrecht Dürer's painting"
"Albrecht Dürer's house"
In the first phrase, the production event implied in "painting" favours this event as link between the two entities. In contrast, in the second phrase, the default possession or ownership relation is more likely to be meant.

5 Conclusion

We presented a partial mapping of CRM event classes to GermaNet, a German wordnet. The mapping is used as a lexicon for detecting event mentions in free text. The mapping does not claim to be complete and will be refined in the future while applied to more textual sources and other cultural heritage domains. Likewise, we will extend the algorithms and tools for event detection so that they better suit the needs of the users.

[22] In fact, determiners have a long history in linguistics of functioning as a discriminator for class or instance.

Acknowledgement

We are very thankful to Guenther Goerz for helpful suggestions and discussions and to the reviewers for valuable hints and suggestions.

References

1. TimeML 1.2.1. a formal specification language for events and temporal expressions (October 2005)
2. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.e.: Definition of the CIDOC Conceptual Reference Model — Version 5.0.4
3. Faruqui, M., Padó, S.: Training and evaluating a german named entity recognizer with semantic generalization. In: Proceedings of KONVENS 2010. Saarbrücken, Germany (2010)
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). pp. 363–370 (2005)
5. Goerz, G., Scholz, M.: Adaptation of NLP techniques to cultural heritage research and documentation. Journal of Computing and Information Technology 18 (2010), http://cit.srce.hr/index.php/CIT/article/view/1918
6. Kunze, C., Lemnitzer, L.: GermaNet - representation, visualization, application. In: Proceedings of LREC 2002. pp. 1485–1491 (2002)
7. Kunze, C., Lemnitzer, L., Lüngen, H., Storrer, A.: Modellierung und Integration von Wortnetzen und Domänenontologien in OWL am Beispiel von GermaNet und TermNet. In: Proceedings of KONVENS 2006. pp. 91–96. University of Konstanz (2006)
8. Pustejovsky, J., Hanks, P., Saurí, R., See, Andrew Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M.: The TIMEBANK Corpus. In: Proceedings of Corpus Linguistics. pp. 647–656 (2003)
9. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: A Robust Event Recognizer for QA Systems. In: Proceedings of HLT/EMNLP. pp. 700–707 (2005)
10. Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines Version 1.2.1 (January 2006), http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf
11. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland (1995)
12. Scholz, M., Goerz, G.: Wisski: A virtual research environment for cultural heritage. In: Raedt, L.D., Bessière, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P.J.F. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 1017–1018. IOS Press (2012)
13. Sennrich, R., Volk, M., Schneider, G.: Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. Hissar, Bulgaria (2013)
14. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of CoNLL-2003. pp. 142–147. Edmonton, Canada (2003)

Mapping ICCD Archaeological Data to CIDOC-CRM: the RA Schema

A. Felicetti[1], T. Scarselli[2], M. L. Mancinelli[3], F. Niccolucci[1]

[1]PIN, Università degli Studi di Firenze, Italy
[2]ICCU, Istituto Centrale per il Catalogo Unico, Italy
[3]ICCD, Istituto Centrale per il Catalogo e la Documentazione, Italy

Abstract. This paper describes the work carried out by PIN (University of Florence) and the MiBAC, in the framework of the ARIADNE project, for mapping the Italian archaeological documentation system to CIDOC-CRM. ARIADNE's primary goal is the implementation of interoperability among archaeological data at a European level, by creating a technological infrastructure for archaeological data sharing and integration. The Italian system is extremely articulated and complex, but the mapping activities, although at an early stage, are progressing very quickly. We are presenting here an overview of the conceptual mapping between the "RA" model (providing information on archaeological artefacts) and CIDOC-CRM, the reference ontology chosen by ARIADNE as a "common language" for integration.

Keywords. Archaeology, Mapping, CIDCOC-CRM, Linked Open Data, Semantic Web

1 Introduction

The activities described in this paper fall within the framework of ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe), an FP7-INFRASTRUCTURES-2012-1 EU project (Grant agreement no: 313193), whose primary goal is the integration of existing archaeological research 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 [1].

Nowadays there is a large availability of archaeological digital datasets, which differ in structure, aims and provided functionalities, representing the outcome of the research of individuals, teams and institutions that altogether span different periods, domains and regions. And since standardization is one of the main keys for integration, it is paramount in this particular moment to find a "common language" for the description of the huge variety of archaeological data available, to make them interoperable and to give the researchers the access to integrated archives for the enhancing of their research activities.

ARIADNE has chosen CIDOC-CRM [2] to implement such integration, and mapping activities have already started within the project to convert data and try to build integrated scenarios. In this paper we present the first phase of the activities carried out by MiBAC, the Italian Ministry of Cultural Heritage and Tourism, and PIN (University of Florence), for mapping the Italian national standards for the encoding of archaeological information, developed and maintained by ICCD, to CIDOC-CRM.

2 The ICCD and the Italian documentation for cultural heritage

The ICCD (Central Institute for Catalogue and Documentation) [3-4] is one of the seven Central Institutes of the MiBAC, whose main goal is to create a centralized national catalogue of the 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 Italian cultural and artistic heritage [5]. Among his most important tasks there is the management of the national general catalogue of archaeological, architectural, historical, artistic and ethno-anthropological heritage, the development of cataloguing methodologies and standards, the coordination of the technical institutions involved in the cataloguing activities on the national territory.

The ICCD is one of the main actors in the realization of the integration between the databases of the MiBAC and the ones of the local institutions distributed on the territory, by means of a number of "regional agreements" with the Regions and the Regional Offices. The Institute promotes dialogue with the territory intended to support the standardization and the integration in the national catalogue, on the basis of the compliance with its cataloguing standards. The agreements also represent the formal approval of a plan of cooperation with institutions put outside the Ministry itself (e.g. dioceses, universities) as part of a systematic action between the Institute and the territorial structures.

The relationship between ICCD and local authorities is fully oriented to the knowledge, the protection and the enhancement of cultural heritage. In this context, the ICCD also provides:

- Standards, methodologies and guides for the technological management of the general catalogue; the cataloguing procedures are monitored and estimated through the ICCD Observatory for Cataloguing (an internal committee in charge of the various management institutions and activities related with the cataloguing activities) [7].
- Tools for data management, and mainly the SIGECweb (Information System General Catalogue), a software web application created with the aim to unify and streamline processes related to the cataloguing activities of the cultural heritage, and to ensure, through the tight control of the applied procedures, the quality of the data produced and their compliance with national standards [8].

3 ICCD Cataloguing Standards

The ICCD *corpus* of cataloguing standards consists of regulations, support and control tools (vocabularies, lists of terms) and a set of rules and guidelines illustrating the methods to be followed for the acquisition and production of cultural heritage documentation [9]. In particular the *corpus* includes:

- *Regulations for cataloguing*, describing the data models and the Authority files [5], to be used for cataloguing activities.
- *Catalogue schemas*: descriptive models and forms for collecting information in an structured way, according to a "path of knowledge". The ICCD issued different cataloguing schemas in relation to different types of assets, organized on the basis of the various disciplines (see below).
- *Authority files*, a complete control system to guarantee uniformity in the use of information concerning key concepts (e.g. authors, bibliography) used throughout the whole system. The Authority files are useful support tools for the standardization of cataloguing, and come as self-consistent databases to be connected with the cultural heritage ones. ICCD created and maintains four Authority files for archaeology, three of which are taken into account for the present paper: "AUT" (Authors), "DSC" (Archaeological Excavations) and "RCG" (Archaeological Surveys).
- *Support and control tools*: thesauri and terminological tools [6] developed to perform data acquisition operation in a uniform way by using similar criteria, and to create a "common and shared language", essential for a correct use of information at query time and for the interoperability of cultural heritage data.

The system is the result of a long research work carried out within the ICCD, in collaboration with other institutions, to develop a model for the acquisition of data that could respond to the needs of a fast cataloguing without compromising a deeper knowledge of the assets. For what concerns the archaeological field, that is the argument of the present paper, the tools available at the moment for the cataloguing of movable and immovable archaeological properties (according to version 3.00 of the Regulation, recently released) are the followings:

- *SI Schema - Archaeological Sites*: used to describe and document an archaeological site, intended as a "portion of land that preserves evidence of human activities, belonging to a past more or less remote, and investigable with the proper methods of archaeological research", with any regard to quality, quantity or size of the evidence.
- *SAS Schema - Stratigraphic Surveys*: used for the documentation of stratigraphic sequences found in contexts of archaeological excavations. The ICCD has an on-going research project for the automatic processing of the records for the detection of Stratigraphic Units, for which, so far, paper forms are the only source available.
- *CA Schema - Archaeological Complexes*, used for the documentation of archaeological properties, without regard of the current state of conservation, having a functional architecture easily identifiable *per se*, both from the physical and conceptual point of view, and composed of various building units (e.g. a fortified place, an *insula*, etc.).
- *MA Schema - Archaeological Monuments*: used for the recording of archaeological properties consisting of a single identifiable building unit (a tower, a *domus*, a temple, etc.), identified and organized on the basis of the functional units (circles) and partitions (walls, roofs, floors, etc.).
- *RA Schema - Archaeological Finds*: used for the recording of movable objects, it is the most used and well established standard for Italian archaeology, because of the very high number of artefacts, already available and continually increasing as a result of archaeological excavations, surveys and discoveries throughout the national territory, and the extremely heterogeneity of types, history and contexts of belonging. For its complexity and completeness, the RA schema is the one we have chosen to start our mapping activities from, as described in this paper.
- *NU Schema - Numismatics*: used for the recording of all the objects mainly having a monetary relevance, not only coins but also object possessing monetary connotation, including seals, ancient medals, coinage tools and weights.
- *TMA Schema - Archaeological Materials*: used for the recording of large collections of materials without significant characteristics or fragmentary, often coming from archaeological excavations or surveys, or stored in museums and private collections, for which it is not expected to use RA schema.
- *AT Schema - Anthropological Finds*: to record biological evidences in close relation with archaeological and paleontological, historical and cultural contexts, affecting the evolution, life and history of studies of the human race and its predecessors.
- *EP Schema - Epigraphic Model:* to record the various aspects of the epigraphic documentation. This model is still under developing.
- *US Schema – Stratigraphic model:* to record the various aspects of archaeological analysis. This model is still under developing.
- *TM Schema – Type wall model:* to record the various aspects of technical wall. This model is still under developing.

The logical organization and interoperability among the various standards listed above provides a comprehensive hierarchic framework for top-down analysis (i.e. from the general 'territorial container' represented by the archaeological site, throughout the archaeological complex, the individual monument composed of parts and subparts, straight to the artefact) and, *vice versa*, to reconstruct the bottom-up sequences from the movable object back to the monumental and territorial context of belonging, according to a strong and articulated system of relationships between the various schemas, which is not rigidly preordained, but can vary to fit different scenarios.

This ICCD reporting system allows, for example, to link archaeological assets of various types to the archaeological site, in which they were found, or to contextualize the stratigraphic investigations in the building, in which they were made (portion of land or monumental emergency), or even to establish correlations between assets of a certain functional or typological kind, to reconstruct funerary objects, collections of objects, sets of artefacts belonging to particular contexts. It is important to note that the whole cognitive process that the cataloguing standards provide is flexible enough to allow the recording of various levels of information, from a minimum number of fields (the so called "inventory level") to a complete and detailed recording of complex data.

To enhance internal interoperability of the system, a parallel work has been carried out to provide all the models listed above with the so-called "cross-sections", special information common to all the models, coming as transversal paths going through the whole system. The cross-sections represent the core, the basic information units around which the specific information and attributes are organized.

4 The ICCD Mapping to CIDOC-CRM

After a deep analysis of the ICCD system, we have agreed that the RA schema is the most significant model of the ICCD archaeological cataloguing system, for its richness and popularity. We have chosen to use it as the starting point for the mapping activities to CIDOC CRM. In facts, RA records contain a huge amount of information for the description of archaeological objects, different types and relationships with other archaeological entities. Moreover, the massive presence in the RA schema of "cross-sections", also present in other schemas, also constitutes a good base for the prosecution of the mapping activities [6-7].

To facilitate the comprehension of the conceptual mapping proposed in this paper, we have chosen to organise the RA information around some of the core concepts of the CIDOC-CRM, in order to give it a semantic order instead of following the functional sequence of descriptions of the RA schema. In facts, although these two sequences coincide in most cases, it is easier to explain the logic of the mapping using a CRM approach, being its model based on events, usually easy to pinpoint and analyse. This is even more necessary in a paper whose main purpose is not to describe in details the whole work carried out, but just to give a general idea of what has been done. Actually, where the words are limited to express such complexity, images could be more effective. For this reason we have tried to synthesize mapping concepts in various figures providing more details. But still, a full description of the whole process remains impossible in this little space.

4.1. Archaeological Object and Identifiers

RA schema concerns the description of artefacts. From the CIDOC-CRM perspective, an artefact is a physical object purposely created by human activity. For this reason, the *E22 Man-Made Object* class has been used for representing the object, which the information in the RA schema refers to.

ICCD records and keeps track of a wide set of identifiers for each object, including the ones inherited by the local institutions contributing to the general catalogue. ICCD also assigns a specific "unique" identifier to the artefact, when it is recorded for the first time in the ICCD archives. In particular, the "NCT" unique code serves as a 'key' to uniquely identify the artefact described in the record at national (Italian) level. It is generated by the combination of various subfields (sub-region code, catalogue numbers assigned by the ICCD). The "NCT" is the most meaningful identifier for the artefact, the one used as the primary and preferred identifier for it. For this reason we used the *P48 has preferred identifier* property of the CIDOC-CRM to relate it with the object.

Another important identifier, deserving to be mentioned here, is the inventory number ("INVN") assigned by the local institutions responsible for the object (i.e. the museum, Superintendence, or private collector holding the property of the artefact).

The "NCT" identifier is already used by ICCD for the creation of uniform identifiers and can be also used in the future for the creation of the persistent URIs for the objects, and for LOD creation and publication.

4.2. Object description

This paragraph provides specific information coming from different sections of the RA schema, describing specific features directly possessed by the artefact and having no direct relation with the CIDOC-CRM events in which the object is involved. In particular ICCD records:

- *Object Definition*: term or expression that identifies the object on the base of its functional and morphological aspect expressed according to the tradition of the studies (e.g. "anfora").
- *Specific Object Typology*: a term referring to the specific class to which the object pertains. This field is usually combined with Object Definition (e.g. "Dressel 20").
- *Production Class*: category, class or type of production to which the object belongs.
- *Object Subject*: the subject or scene represented by the object (only for objects that represent themselves an iconographic subject).

ICCD provides specific vocabularies for the definition of the typological fields described above. All of them have been mapped on the *E55 Type* class, and linked to the archaeological object via the *P2 has type* property.

Other features directly referring to the object are:

- *Object Name*: the historical or traditional name of the object or its dedication name (e.g. "Olpe Chigi"). It corresponds to the *E35 Title* class.
- *Position*: this field represents a very peculiar case, since it indicates the name of the current object with respect to a larger object of which is part (e.g. "foot", saying that the current object is a foot of, for instance, a statue). To render this concept, we have used the *E46 Section Definition* class and the related properties.
- *Title*: the title given by the author or the traditional name given to the object (i.e. "Apollo del Belvedere"), mapped on the *E35 Title* class.
- *Materials*: materials of which the object is made, described using the *P45 consists of* property and the *E57 Material* class. A specific vocabulary is provided.
- *Dimensions*: information concerning the various dimensions of the object (e.g. height, width, length, etc.), including the estimated monetary value of the object calculated on the currency at recording time. The *E54 Dimension* and the related properties (*P43, P90, P91*) have been used for the mapping of these fields.
- *Features carried by the object*: inscriptions (dedicatory, commemorative, honorary, etc.), stamps, badges, emblems and other features indicating e.g. the original property or provenance of the object. The RA schema devotes a special section to the description of these objects and their characteristics. For inscriptions, in particular, it records language, transcription, character set, writing technique and the cultural area of be-

longing (e.g. Roman or Greek epigraphy). CIDOC-CRM is particularly suitable for describing inscriptions and provides a complete set of entities and properties for it (i.e. the *E34 Inscription* class and the related properties) [8].

- *Physical conditions and state of preservation of the object*. We used the *E3 Condition State* together with the *P44 has condition* for the mapping, and the E55 class to record the terms of the controlled vocabulary provided by ICCD for populating this field.
- *Information on digital items*, such as pictures, drawings, multimedia, etc., documenting the object. The CRM *E36 Visual Item* and *E38 Image,* together with the *P138 has representation* property, have been used for mapping these fields.

4.3. Locations and Places

RA includes specific sections ("LC" and "LA") for the description of the various locations where the object was produced or found, where is currently located or was located in the past. The terminology for the definition of these locations is based on the UNI-ISO 3166-1 standard (alphabetical list of country names) and on the standard lists of terms for the Italian administrative areas (regions, provinces and so on) provided by ISTAT (the Italian Institute for Statistics). The indication of all places on the Italian territory follows the ICCD standard path "Region > Province > Municipality > Locality". For the purposes of the current mapping, these information could be easily enriched with GeoNames URIs, to enhance future interoperability (see Figure 1).

A list of the different location types recorded into the archive, with indications on how they were mapped to CIDOC-CRM, follows.

- *Current location*: is described in section "LC" (*Geo-Administrative Location*) with a set of fields providing identification of the geographic and administrative place on the Italian territory or to administrative-territorial organizations of foreign countries (in the case, for example, of objects held in areas pertaining to the Italian embassies) where the artefact was located at the moment of the ICCD record creation. To map the notion of "current location" to CIDOC-CRM, we linked the instance of *E53 Place* with the archaeological object through the *P55 has current location* property.
- *Provenance places*: described in section "LA" (*Other Geo-Administrative Locations*), it provides information not only for the geo-political localizations of the object's previous places of conservation, but also for production and finding places, according with the "TCL" field (*Location Type*) whose value (*Provenance*, *Finding*, *Production*) determines the mapping to be followed. When the section refers to the object provenance, all the fields are assumed to be repeatable. This is very useful for the reconstruction of the object's location history, i.e. the sequence of all the places in which it was present through time. CIDOC-CRM is very handy for this, since it also gives the possibility to define events able to relate places, actors and time spans, even if they are described in different sections of the original data schema. In this case, to relate the object with one of its provenance places, we have created the *E10 Transfer of Custody* event and specified the provenance place by using the *P7 took place* property. The object participation in this event is defined via the *P30 custody transferred through* property.
- *Production* and *Finding Place:* the information of section "LA" refers to the corresponding place type with "TLC = Production" or "TCL = Finding". Details on these place types are provided in the "Production" and "Finding" paragraphs of this paper.

Figure 1 illustrates the general mapping schema of ICCD locations and places.

Information concerning each place described in the archive includes:

- *Specification of the architectonic or functional typology* of the place or building in which the object is currently located or /was located in the past (e.g. "Museum", "Abbey", "Monastery"). ICCD provides typological thesauri for these fields, which can easily be mapped using the *P2 has type* property, and assigned to the specific *E55 Type*.

- *Denomination*, i.e. the full name of the place, building or complex where the object is currently hosted or /was hosted in the past. For the name of buildings ICCD makes reference, where possible, to official sources (e.g. the "Diocesan Yearbooks" for church buildings). The *E44 Place Appellation* is used to assign denominations to places. The *P89 falls within* property is used for stating the mutual relationships among different places (e.g. between a building in respect to the complex it belongs to).
- *Addresses*, mapped on the *E45 Address* entity.
- *Denomination of the collection* which the object forms (or formed) part of (*P46*), hosted in a specific place (*P55 has current location*).
- *Related date*, i.e. the date on which the object was placed in the museum/building (*P26*) and the one in which he was transferred elsewhere (*P27*).
- *Spatial coordinates* (*E47*) defining the points needed to identify (*P87*) and georeference the place where the object was held or is currently located. Spatial coordinates also refer to all the other place types described by ICCD (see below). Information on specific techniques and methods used to acquire the coordinates are also provided.

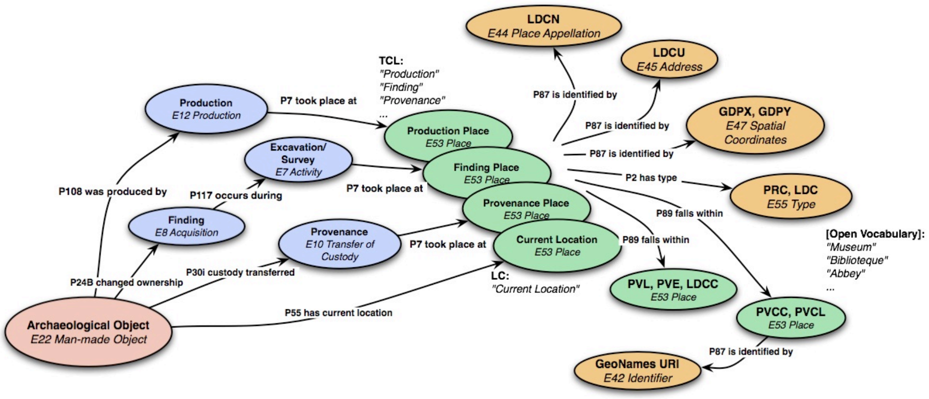


Fig. 1. CIDOC-CRM mapping of the ICCD-RA locations and places information

4.4. Finding

The finding event is a very important event in archaeology, representing a corner stone in the reconstruction of the object's history. From the CIDOC-CRM point of view, the object finding is a kind of acquisition (*E8*) that can occur during (P117) an archaeological survey or excavation (E7) and changes the object's ownership (*P24B*), which is acquired by the institution performing the discovery. The ICCD RA schema provides, in the "RE" section, a wide bunch of information concerning finding activities, and in particular:

- ICCD unique identifier (through the "DSC" Authority file) and Excavation inventory number (*E42*).
- Official name and description of the archaeological excavation/survey, mapped as instances of *E41 Appellation*.
- Information concerning institutions, scientific coordinators and other people responsible or involved in the survey/excavation, during which the object was found. Each of them has been mapped as an instance of *E39 Actor.*
- Survey/excavation motivations (*P17*, e.g. "Rescue archaeology").

- Methods and techniques (*P32 -> E55*) used to perform the excavation/survey activities. Terms to specify this field are taken from a specific vocabulary.
- Time spans (*P4 → E52*)
- Specific bibliography, documenting (*P70*) the finding activities.
- Finding places: a set of fields providing information on the place where the object was found (section "LA" with "TCL = Finding")
- Stratigraphic units, tombs and other locations where the finding took place (*P7*). Figure 2 shows the mapping rationale of the "Finding" event and the related entities.

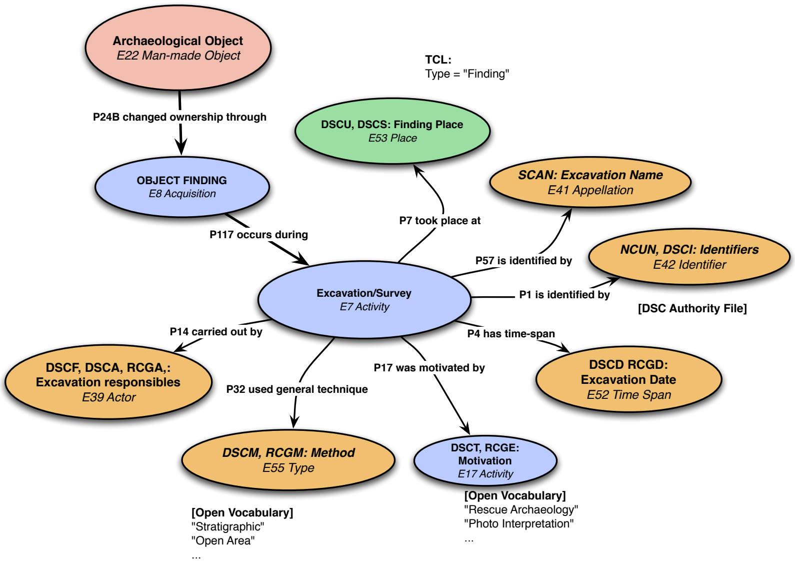


Fig. 2. Mapping schema of the ICCD-RA "Finding" event.

4.5 Production

Production is a very complex process, involving various objects, people and places. Documenting it in the proper way is paramount when dealing with archaeological artefacts. In a similar way to what we have done for the finding, we have defined a production event (*E12*) able to relate each other the various places and actors involved. The archaeological object's production is specifically referenced by the *P108 was produced by* property.

Production is described using data coming from various sections of the ICCD schema, in which we find all the information to describe the creators and the techniques involved in the object production process (*P32*), but also notices about the group of artists or the school and other similar concepts related to a more general cultural context. ICCD, as already mentioned, defines a specific Authority file for the "authors" ("AUT"), providing unique identifiers to be used here for the unambiguous identification of all the actors participating in the production process. We used the *P14 carried out by* to relate these actors with the production event (*E12*).

ICCD also records information concerning the reasons for the attribution of the object to a certain cultural context. We have rendered this attribution process by using the *P140 was attributed by* property and the *E13 Attribute Assignment* event. ICCD provides a controlled list of terms for production attributions, which we used to define the attribution type (*E55*, e.g. "stylistic analysis").

The schema also contains fields providing specific information on production place, if known.

4.6. Acquisition

The ICCD *Acquisition* section ("ACQ") records information related with the acquisition and the legal *status* of the artefact, the protective measures concerning it and information related with the circumstances under which the object has been received and is located in the current conditions of property or detention. Since institutions may have various ways for acquiring an archaeological object (e.g. after a finding or by a purchase, a donation, an exchange, etc.), ICCD has specific vocabularies for the definition of acquisition types (mapped on *P2*). The Acquisition event (*E8*) in section "ACQ" is considered as the changing of ownership of the artefact through a transfer of title from a former owner (*P23*) to a new one receiving its ownership (*P22*). The section can appear many times to document the acquisition chain occurred during the object's lifetime.

The "ACQ" section also records the acquisition dates and places, and provides details concerning the actors (people or groups) involved in the event. For the latters, the *P52 has current owner* is used to define the last recipient, in our case the institution that created the record (*E39*).

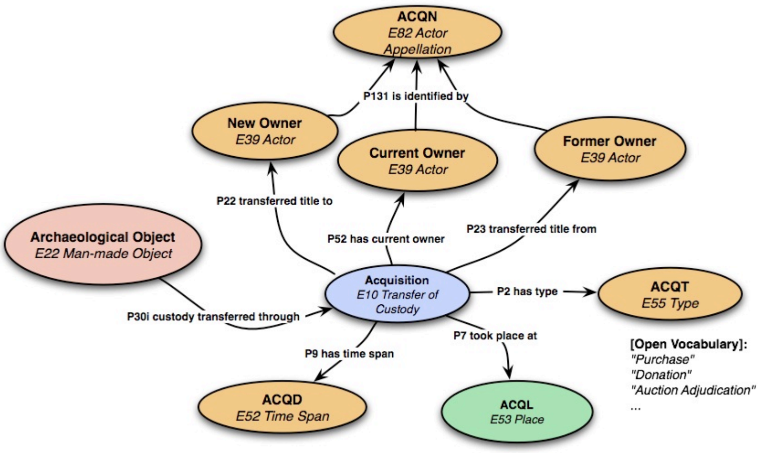


Fig. 3. The Acquisition process and the "ACQ" section of the RA model.

4.7. Objects dating

Dates are usually very problematic information to manage, for their notorious imprecision which always makes it impossible to record them in a standard way. In ICCD RA model, we find various chronological indications for dating the objects, including periods of reference (e.g. "Middle Neolithic"), centuries in Roman numerals, numeric expressions followed by the indications a.C. (BC) or d.C. (AD) (e.g. "III sec. a.C."), and chrono-cultural definitions (e.g. "Roman Age").

Since historical periods do not possess universally agreed start and end chronological limits, we have used the *E4 Period* entity to represent them and the *P10 falls within* to establish relationships with the object production event (*E12 Production*). Sometimes ICCD provides a single date or *termini ante* and *post quem* for the definition of data ranges. In this case the *P82 at some time within* property has been used.

As in the case of the reasons for the attribution of production to a certain cultural context, ICCD provides information concerning the reason for the proposed dating of the object (the "DTM" field, *dating motivation*). In a similar way to the above, we have rendered the dating attribution process by using the *P140 was attributed by* property and the

E13 Attribute Assignment event. A controlled list of terms for dating motivations, used to define the attribution type (*E55*, e.g. “chemical analysis”), is also provided.

4.8. Internal ICCD cross references

A special section of the RA schema states direct relationships between entities catalogued using the ICCD system. These relations are only specified if both the objects (the one described in the current record and the one referenced from this section) are present into the ICCD database. The field “RSER”, in particular, defines the nature of the relationships described in this section and as a consequence, from the point of view of the mapping, the path that should be followed according with it. The same field can also determine the place type (i.e. the current location, the provenance, the finding or the production place, similarly to what “TCL” field does) involved in the relation. The “RSEC” and “RSET” fields indicate the object type of the referenced asset and its unique identifier.

4.9 ICCD RA Bibliography

The ICCD, while not providing the completeness typical of library databases, records a well-detailed bibliographic information set concerning the archaeological objects. In the mapping, the object is linked via the *P70 is documented in* property with its bibliographic record (*E31 Document*), which in turn has been created by an event (*E65*) having specific actors and creation dates. The “AUT” Authority file is used for authors’ definition. The *P3 has note* property has been used to assign the full citation to the document itself.

5 Mapping Example

In the following table we propose a real example of mapping of an artefact (Olpe ‘08-487640’), found in 1969 during the excavation of the archaeological area of Sasso Marconi and exhibited in the Etruscan National Museum of Marzabotto (Bologna, Italy). Details of the mapping paths (column 2) and of the ICCD vocabularies used (column 3), are also provided.

ICCD RA	CIDOC-CRM Mapping	Vocabularies
NCT - Codice univoco	E22 Man-Made Object - P48 has preferred identifier - E42 Identifier *‘08-487640’*	
OGTD - Definizione	E22 Man-Made Object - P2 has type - E55 Type *‘olpe’*	Open Vocabulary
CLS - Classe e produzione	E22 Man-Made Object - P2 has type - E55 Type *‘contenitori e recipinenti’*	Categories
PVCC - Comune	E22 Man-Made Object - P55 has current location - E53 Place P89 falls within - E53 Place *‘Marzabotto’*	ISTAT Names
LDCT - Tipologia	E22 Man-Made Object - P55 has current location - E53 Place P2 has type - E55 Type *‘museo nazionale’*	Open Vocabulary
LDCN - Denominazione	E22 Man-Made Object - P55 has current location - E53 Place P87 is identified by - E44 Place Appellation *‘Museo Nazionale Etrusco «Pompeo Aria»’*	
INVN - Numero	E22 Man-Made Object - P1 is identified by - E42 Identifier *‘1437’*	
SCAN - Denominazione dello scavo	E22 Man-Made Object - P24B ch. own. thr. - FINDING (E7 Activity) P57 is identified by - E41 Appellation *‘Sasso Marconi, Via Porrettana 252/3’*	

DSCF - Ente responsabile	FINDING - P14 carried out by - E39 Actor *'SBA-ERO'*	
DSCT - Motivo	FINDING - P17 was motivated by - E7 Activity *'opere private'*	Open Vocabulary
DSCD - Data	FINDING - P4 has time-span - E52 Time Span *'1969'*	
MTC/M - Materia	E22 Man-Made Object - P45 consists of - E57 Material *'bronzo'*	Open Vocabulary
MTC/T - Tecnica	E22 Man-Made Object - P108 was produced by - E12 Production P32 used technique - E55 Type *'laminatura, fusione'*	Open Vocabulary
DTZG - Fascia cronologica di riferimento	E22 Man-Made Object - P108 was produced by - E12 Production P10 falls within - E4 Period *'sec. V a.C.'*	
DTM - Motivazione cronologia	E22 Man-Made Object - P108 was produced by - E12 Production P10 falls within - E4 Period P140 was attributed by - E13 Attribute Assignement *'contesto'* *'analisi tipologica'*	Open Vocabulary
MISU - Unità	E22 Man-Made Object - P43 has dimension - E54 Dimension P91 has unit - E58 Measurement Unit *'cm'*	Closed Vocabulary
MISA - Altezza MISD - Diametro	E22 Man-Made Object - P43 has dimension - E54 Dimension P90 has value - E60 Number *'18,4'* *'8,8'*	
DESO - Indicazioni sull'oggetto	E22 Man-Made Object - P3 has note *'Bocca rotonda, labbro estroflesso, brevissimo collo troncoconico, corpo globulare senza soluzione di continuità con il fondo etc.'*	
STCC - Stato di conser.	E22 Man-Made Object - P44 has condition - E3 Condition State *'reintegrato'*	Closed Vocabulary
ACQT - Tipo acquisizione	E22 Man-Made Object - P24 ch. own. thr. - E10 Transfer of Custody P2 has type - E55 Type *'scavo'*	Open Vocabulary
ACQD - Data acquisizione	E22 Man-Made Object - P24 ch. own. thr. - E10 Transfer of Custody P9 has time span - E52 Time Span *'1969'*	

6 Data Conversion and Publication

The conceptual mapping described in this paper is the logical base on which data encoded with the RA model can be converted in a CIDOC-CRM RDF format. Implementation of the real data conversion can be performed in various ways, but of course, the most suitable one would be using the exporting features already provided by SIGECWeb. The official ICCD software, in facts, is already able to export information concerning entities, cross-references and internal relationships, in various ways. The preferred and most used one is the ICCD "exporting package", mainly a set of directories containing textual data descriptions and multimedia files. Since the textual information always remains compliant with the various ICCD models, implementing the mapping framework and converting it directly in RDF is very straightforward. The system is also able to export data in XML, which would further simplify the converting operations and the generation of semantic data in Linked Open Data format.

Anyway, the ideal scenario would be reached by implementing new SIGECWeb modules and facilities for the direct CIDOC-CRM RDF exporting, and the direct publication of semantic information as Linked Open Data on the institutional websites of the MiBAC. This would simplify the conversion operations and constitute a tremendous step forward on the road of the interoperability of cultural heritage information. Publication would also be straightforward, since the MiBAC online infrastructure already provides many RDF frameworks for the hosting and management of semantic information, together with various SKOS and Linked Open Data facilities for the semantic web implementation [9].

7 Conclusions and future work

The RA Schema is only the beginning of a wide activity that will be carried out by the ARIADNE project in collaboration with ICCD and other institutions related with MiBAC. The mapping of this complex schema has already demonstrated, at least from the logical point of view, the coherence with CIDOC-CRM and a wide compatibility with its schema. Though, a lot of work remains to be done. ICCD is still completing its model, and a version 4.00 of the recommendations for cataloguing, making it even more rational, is going to be released. From the other side, CIDOC-CRM is also evolving and an extension specifically designed to capture the concepts of the archaeological field is going to be released as part of the ARIADNE outcomes. The RA mapping will surely constitute a good starting point for the future convergence of the two models. And, on top of it, common concepts and elements like the cross-sections will make the mapping of all the other ICCD archaeological schemas easy and fast to be performed.

ARIADNE will assist ICCD in building and evaluating this process in every phase, from logical mapping to physical conversion of archaeological data. ARIADNE is also carrying out similar activities with other European archaeological institutions (partners of the project) to achieve, in a near future, its main goal: the implementation of interoperability among archaeological data at a European level.

References

1. ARIADNE Project, http://www.ariadne-infrastructure.eu/
2. CIDOC-CRM, http://www.cidoc-crm.org
3. ICCD, http://www.iccd.beniculturali.it/index.php?it/1/home
4. Le Boeuf, P., Doerr, M., Ore, C. E., Stead, S.: Definition of the CIDOC Conceptual Reference Model (2012). http://www.cidoc-crm.org/official_release_cidoc.html
5. Ministero per i Beni e le Attività Culturali, Istituto Centrale per il Catalogo e la Documentazione, Rapporti. 1 (2001), http://www.iccd.beniculturali.it/getFile.php?id =409; 2 (2003), http://www.iccd.beniculturali.it/getFile.php?id=410; 3 (2007), http: //www.iccd.beniculturali.it/getFile.php?id=411; 4 (2009), http://www.iccd.benicultu rali.it/getFile.php?id=412
6. Eide, O., Felicetti, A., Ore, C. E., D'Andrea, A., Holmen, J.: Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC-CRM Mapping. In: EPOCH Conference on Open Digital Cultural Heritage Systems, pp. 1-7. (2008). http://public-repository.epoch-net.org/rome/05%20Procedures%20Data%20Integration.pdf
7. D'Andrea, A., Marchese, G., Zoppi, T.: Ontological Modelling for Archaeological Data. In: VAST 2006, The evolution of Information and Communication Technology in Cultural Heritage, pp. 211–218, Budapest, Archaeolingua (2006). https://diglib.eg.org/EG/DL/WS/VAST/VAST06/211-218.pdf.abstract.pdf
8. Doerr, M., Dionissiadou, I.: Data Example of the CIDOC Reference Model - Epitaphios GE34604. Benaki Museum, Athens (1998). http://www.cidoc-crm.org/docs/crm_example_1.pdf
9. Felicetti, A.: MAD: Managing Archaeological Data. In: VAST 2006, The evolution of Information and Communication Technology in Cultural Heritage, pp. 124–131, Budapest, Archaeolingua (2006). http://public-repository.epoch-net.org/publica tions/VAST2006/project1.pdf

Pattern based mapping and extraction via CIDOC CRM

Douglas Tudhope[1], Ceri Binding[1], Keith May[2], Michael Charno[3]
([1]University of South Wales, [2]English Heritage, [3]Archaeology Data Service)

douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk
keith.may@english-heritage.org.uk
michael.charno@york.ac.uk

1 Introduction

The current situation within archaeology is one of fragmented datasets and applications, with different terminology systems. The interpretation of a find may not employ the same terms as the underlying dataset. Searchers from different perspectives may not use the same terminology. Separate datasets employ distinct schema for semantically equivalent information. Entities and relationships may have different names but be semantically equivalent. Even when datasets are made available on the Web, effective cross search is hampered by semantic interoperability issues [1].

It is becoming increasingly understood that the use of an integrating conceptual framework, such as the CIDOC Conceptual Reference Model (CRM) (ISO 21127:2006) [2, 21], can help address these issues. We take this as our agreed point of departure. This paper discusses various implementation issues to facilitate use of the CRM. Employing the CRM has tended to require an understanding of the source dataset schema and also specialist knowledge of the CRM and techniques for mappings. This paper argues for the use of mapping patterns to guide deployment, to improve homogeneity, to increase data interchange and to encourage greater uptake.

1.1 Relevance to CRMEX Workshop

This paper discusses our implementation experience related to the issues raised in the call for papers of the CRMEX Workshop:

- Because CRM allows many different ways of representing the same situation, CRM adopters in various cultural heritage areas need mapping guidelines and best practices to increase the chance of interoperation.
- While Resource Description Framework (RDF) is a viable CRM representation, there are various low level RDF issues that are not standardized. Since RDF representation implies a certain implementation bias and still undergoes changes of good practice, the CRM Special Interest Group (CRM-SIG) has been expecting good practices to emerge from people applying CRM in order to make recommendations.

The work presented here discusses experience with our development of lightweight techniques and tools to map and extract CRM-based archaeological data with final publication as Linked Data. These techniques have been used in significant CRM-based implementations in two projects STAR [6] and STELLAR [7] described below.

At the Workshop on the Implementation of CIDOC-CRM, organised by the German Archaeological Institute (DAI) in Berlin 2009 [8], we raised the following CRM implementation issues from our experience in the STAR project:

- For application interoperability we need agreement on lower level implementation representations (e.g. data types, date formats, spatial coordinates etc.)
- Need provision of vocabulary (terminology) - our approach is to employ SKOS to model vocabulary elements and link to CRM [19]
- CRM can be extended for domain specificity
- CRM is event-based and therefore
  - Mapping a data property to CRM typically results in a chain of CRM relationships
  - Directly representing the model results in complex user interfaces
  - There is a need for user interface ‘short cuts’ and simplified views for particular purposes
- Data can be mapped to multiple CRM elements depending on what is considered relevant and important - need for guidelines as to the focus and purpose of a mapping exercise

We next describe briefly the STAR and STELLAR projects, where we explored the above issues. This paper focuses mainly on a discussion of mapping issues (details of our implementations are given elsewhere but we are happy to discuss in the workshop). We then consider issues raised at the 2009 DAI workshop, together with a discussion of the pattern based approach we have adopted as one way of addressing the issues.

2 STAR Project

The STAR (Semantic Technologies for Archaeological Resources) project was a collaboration between the Hypermedia Research Unit at the University of South Wales (formerly Glamorgan) and English Heritage (EH). The project aimed to provide a degree of semantic interoperability between diverse archaeological datasets from different projects and organisations. The system makes cross-search possible on excavation datasets including Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman and Stanwick sampling together with archaeological reports extracted from the OASIS grey literature library, provided by the Archaeology Data Service [9].

Since the CRM operates at a relatively high level of generality, the datasets were mapped to the CRM-EH archaeological extension of the CRM, developed by English Heritage [3, 4]. For working with archaeological datasets at a more detailed level, the CRM-EH specializes the CRM classes for Physical Object and Place to archaeological subclasses such as Find and Context. In collaboration with EH, an RDF implementation was created [4], referencing and complementing the existing published (v4.2) RDFS implementation of the CRM [5].

Domain expert May generated a series of spreadsheets showing the key mappings from the various datasets to the CRM-EH. Selections from the different databases were extracted via SQL queries; and converted to RDF using a data extraction and conversion tool [10].

Despite the use of the data extraction tool the exercise proved time consuming. The initial mappings produced were incomplete and under-specified, relating selected data fields to CRM-EH entities but often at a higher level than that required for implementation. The fully formed intermediate chains of events and relationships necessary for connecting the entities together had to be deduced in each case and conventions unilaterally decided for important implementation details, such as formats for identifiers, coordinates and measurement units.

The online STAR demonstrator cross searches excavation datasets from the five different databases, together with metadata representing an extract of excavation reports from the OASIS grey literature library [22]. STAR did not necessarily seek to represent each dataset in its entirety but focused on specific inter-site cross search use cases. Previously cross search was not possible; each dataset remained in its own silo, and no link was made to grey literature. The demonstrator seeks via the user interface to hide the complexity of the underlying ontology, while offering structured semantic search. An interactive query builder offers search (and browsing) for key archaeological concepts such as Samples, Finds, Contexts or interpretive Groups with their properties and relationships. As the user selects via the interface, an underlying semantic query is automatically constructed in terms of the corresponding ontological model.

STAR employed a web service architecture for programmatic access to the data and to various glossaries and thesauri. The latter were represented in the W3C standard Simple Knowledge Organization System (SKOS) format [11], a formal RDF representation. EH thesauri were available for programmatic access via a web service API, with extensions for semantic concept expansion [20]. The web services were accompanied by a variety of 'widget' controls that could be integrated into browser based user interfaces, where browsing of concept structures or concept based search is required. In more recent work, we have published national heritage thesauri as Linked Data [12].

Natural language processing information extraction techniques were applied to identify key concepts in the grey literature, producing semantic metadata in the same CRM-EH based representation as the extracted data. This metadata allowed unified searching of the different datasets and the grey literature in terms of the semantic structure of the CRM-EH ontology [23].

The CRM and CRM-EH do not supply a vocabulary of concepts beyond the class names in the ontology. Therefore a selection of thesauri and glossaries were used in conjunction with the ontology for search purposes. An extended set of EH glossaries were closely identified with associated fields in the datasets. This required an intellectual alignment operation to cleanse and align the data with controlled vocabulary concept identifiers – an important aspect of the work. These vocabularies afforded semantic search in the demonstrator, with controlled terms being interactively suggested by the query builder.

2.1 STELLAR

STAR served as the launching point for STELLAR (Semantic Technologies Enhancing Links and Linked data for Archaeological Resources) [7], a collaboration between the University of South Wales and the Archaeology Data Service, with EH as Project Partners. We addressed the mapping difficulties discussed in Section 2 by developing new STELLAR tools to make the process more standardised and to facilitate use by third-party data providers. The aim was to make it easier for data owners who are not ontology specialists to express their data in terms of the CRM (and CRM- EH) and to generate Linked Data representations. The STELLAR tools convert archaeological data to RDF in a consistent manner without requiring detailed knowledge of the underlying ontology.

These tools work from a set of templates that express commonly occurring patterns encountered in the STAR project. A set of pre-defined templates is provided but user-defined templates can also be created. The current set of templates corresponds to the general aim of cross-searching excavation datasets for inter-site analysis and comparison. Different templates drawing on other areas of the ontology (and the datasets) could be designed for purposes such as project management and workflow or detailed intra-site analysis. Each template input is a combination of various optional fields with a mandatory ID. The ID is prefixed with a namespace (supplied by the user) to generate URIs. Thus the RDF output is produced in a form that facilitates subsequent expression as Linked Data. The STELLAR template-based method can be considered as a form of the *pattern based approach* that has recently emerged within Linked Data generally [18].

In addition to CRM-based templates, there is a template allowing a glossary or thesaurus connected with the dataset to be expressed in SKOS. The CRM templates have fields giving the (preferred) option of expressing controlled data items as URIs (either to local vocabularies generated by the SKOS template, or to external Linked Data URIs).

Figure 1 is an example of a pattern to model the relationships between an object, a production event and a material.

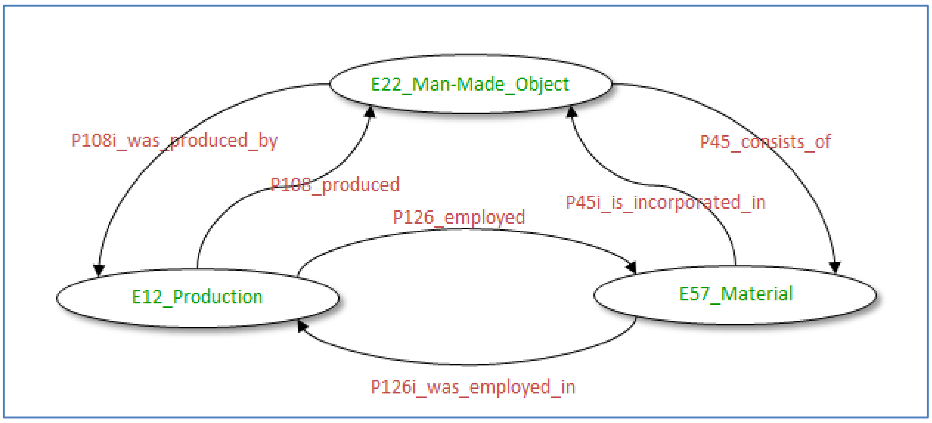


Figure 1. Example pattern

In Figure 2 we see (an extract of) input to the template and then the template itself, which creates directional relationships, an event based property and a shortcut. The user needs to select the particular template (e.g. from a template library) as appropriate for the pattern they wish to express and then supply the data from their datasets. The template contains placeholders corresponding to named columns in the input.

id	material
123	copper

```
  // HEADER template, is output once at start
  HEADER(options) ::= <<

       <?xml version="1.0"?>
         <rdf:RDF
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:crm="http://www.cidoc-crm.org/cidoc-crm/">

  >>
  // end of HEADER template

  // RECORD template, is output once per data row
  RECORD(options, data) ::= <<

  <crm:E22_Man-Made_Object             rdf:about="http://myexam-
ple/E22_$data.id$" />
  <crm:E12_Production rdf:about="http://myexample/E12_$data.id$"
/>
  <crm:E57_Material  rdf:about="http://myexample/E57_$data.mate-
rial$" />
```

```
  <rdf:Description rdf:about="http://myexample/E22_$data.id$">
  <crm:P45_consists_of            rdf:resource="http://myexam-
ple/E57_$data.material$" />
  <crm:P108i_was_produced_by      rdf:resource="http://myexam-
ple/E12_$data.id$" />
  </rdf:Description>

  <rdf:Description   rdf:about="http://myexample/E57_$data.mate-
rial$">
  <crm:P45i_is_incorporated_in    rdf:resource="http://myexam-
ple/E22_$data.id$" />
  <crm:P126i_was_employed_in      rdf:resource="http://myexam-
ple/E12_$data.id$" />
  </rdf:Description>

  <rdf:Description rdf:about="http://myexample/E12_$data.id$">
  <crm:P108_has_produced          rdf:resource="http://myexam-
ple/E22_$data.id$" />
  <crm:P126_employed              rdf:resource="http://myexam-
ple/E57_$data.material$" />
  </rdf:Description>

  >>
  // end of RECORD template

  // FOOTER template, is output once at end
  FOOTER(options) ::== <<
      </rdf:RDF>
  >>
  // end of FOOTER template
```

Figure 2. Example of a STELLAR template and input extract

Templates are available from the STELLAR website, along with tools that operate over the templates. To generate RDF, the user chooses a template for a particular data pattern and supplies the corresponding input from their database. Documentation and a tutorial are available on the website [7]. The Archaeology Data Service used the STELLAR tools to publish Linked Data from a (new) selection of their archived excavation datasets [13].

3 CRM implementation experience from 2009 DAI workshop

Two other projects at the 2009 DAI workshop raised overlapping issues though following different specific implementation methods. The CLAROS project [14] followed a pattern based approach by requiring data providers to conform to a set of XML format CRM patterns [15]. The BRICKS project discussed below encountered various problematic issues when attempting semantic interoperability via the CRM.

The BRICKS FP6 IP project [16] employed spreadsheets to intellectually define mappings from two different archaeological databases to the CIDOC CRM. These were semi-automatically transformed to XSL style sheets, which transformed the data to the desired representation. They experienced consistency problems which resulted in different mappings for the same underlying semantics and in different data objects being mapped to the same CRM entity. They suggested a need for additional technical specifications for implementation modeling purposes. The abstractness of the CRM and the lengthy relationship chains arising from the event-based model also raised issues for designing appropriate user interfaces.

Further details are elaborated in [17] with various potential opportunities for divergent mappings of the same semantics outlined. Examples are given below (Figure 2 illustrates the first two points):-

- Should an E57 Material (e.g. *gold)* be mapped as a property of an E11 Modification event or as a property of an E22 Man-Made Object?
- Should a method of manufacture (e.g. *hammered*) be mapped as an E55 Type of an E12 Production event or as an Appellation of an E29 Design or Procedure?

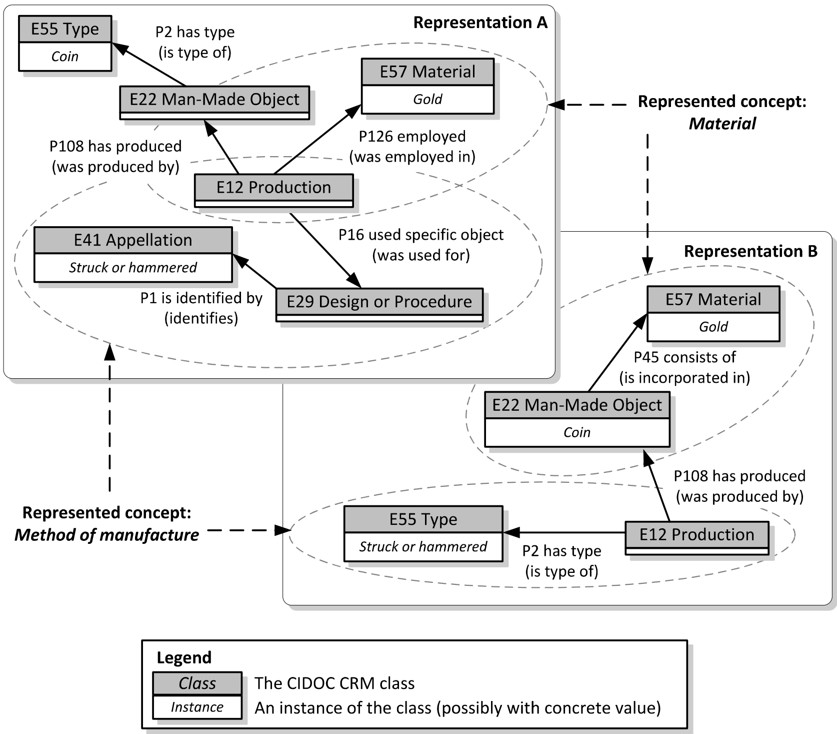


Figure 4: Different valid CRM representations for equal metadata attributes

Figure 2 – a figure taken from [17] illustrating the previous points

Note that the alternatives in Figure 2 are not necessarily equivalent; using a material does not necessarily mean incorporating it in the product and being incorporated does not always imply its use in production. For this instance, both mappings were seen as equally possible in [17] – the note associated with the coin reads “Roman Gold aureus of Nero (AD 54-68) …”. Their argument is for more guidance on defining the mapping paths.

- Should E22 Man-Made Objects be directly identified by an E42 Identifier or should the connection be made via a record that has an Identifier? Due to the CRM’s origins in museum documentation systems, CRM-based integration work has sometimes modeled the record of an object as an entity in its own right. This can give rise to differences with approaches that seek to directly model an object without noting any existing catalogue or recording element.

- All CRM classes can be assigned types (used for domain terminology). This allows different judgments as to whether a thesaurus or gazetteer element should be associated with an object or related activity (or indeed any property).

In addition to the various mapping choices outlined above we can also note that core ontologies offer the flexibility of capturing different aspects of an object, depending on intellectual judgment. Depending on the end purpose of the mapping exercise, a given aspect may or may not be important to model, as for example perhaps with Man Made Objects and Legal Objects, or man-made features. This will naturally vary between different collections with different areas of focus.

Since the CRM is event-based, the issue of when it is appropriate to create an assignment event when assigning an attribute to an object is ever present. Essentially this depends whether the decision to assign an attribute is considered worthy to record. Is the time and actor involved important? Might others judge differently now or in the future? Again this can result in different mapping expressions depending on the judgement.

It could be argued that the choice to model either a shortcut property or a longer fully formed event-based chain adds flexibility. However, inevitable inconsistencies of approach can result. The STELLAR solution is for the templates to automatically generate a pattern of entities and properties consistently modelling *both* possible approaches simultaneously, thus reducing inconsistencies and the requirements for end applications to detect or predict which particular modelling approach has been taken.

Different mappings can potentially pose significant problems for semantic interoperability. It indeed proved a problem for the BRICKS project, which required the addition of an intermediate mapping which itself served as the integrating layer rather than the CRM. In fact, any general core ontology will permit various mappings from the same set of data elements depending on end purpose and focus.

In principle, end-application systems, capable of intelligently traversing the different CRM graphs produced by differences in mapping practice and differences in the granularity of detail and events modelled, could automatically address the issue of different mappings. In previous work with the Art and Architecture Thesaurus, we have implemented faceted query expansion [24]. With regard to the CRM, Tzompanaki and

Doerr [25] discuss the potential for automatic reasoners to take advantage of transitive properties, propagating down from a query expressed in terms of small set of high level fundamental categories and properties (or offering successive specialised choices to the user). While this offers potential approaches for starting from high level facets, in some use cases the ability to start from lower level query patterns is desirable. The performance issues remain to be fully explored (they point out the deficiencies of SPARQL for such complex queries).

The potential to employ reasoning over the CRM graph is indeed one of the reasons for semantic integration. It defeats the point of integration if everyone must say exactly the same thing with the CRM! Nonetheless in our view, a multiplicity of approaches for similar data will pose unnecessary problems for implementation in the medium term. It is not clear that all the problems described by the BRICKS team could be solved by transitive closure alone. Specific rules will probably be required, which raises difficulties for generalising and introducing a new alternative mapping. A pragmatic approach is to combine developments in reasoning with efforts at consensus on patterns for CRM mappings and guidelines. This could involve patterns for particular domains and also general patterns for common situations.

4 Conclusions

When the CRM was originally created the practical context for automated cross search was more limited and it was in part an intellectual resource. Today there is an expectation that any integrating ontology will be employed in machine readable form for automatic semantic interoperability purposes. However, if different implementations of the CRM follow different low level implementation specifications or employ different mappings for the same underlying semantics then this raises barriers for semantic interoperability.

Issues with mapping are probably inevitable in a general ontology intended to capture a wide range of practice and, as with the application of general library classification schemes, different choices for realising a collection in the CRM may be expected. However the potential divergence of mapping practice poses challenges for implementations and the final applications, particularly where it cannot be assumed that such applications possess built in reasoning capabilities that could ameliorate some of the differences.

Thus the purpose (or use case) of any shared mapping exercise should be stated if possible. Data providers or those responsible for mappings should have available (if they choose) *mapping patterns* and corresponding guidelines for their domain or the mapping exercise in question.

Working from established RDF patterns guarantees the semantic interoperability of the resultant data and also that the syntactical implementation details are handled consistently. It is also more friendly to non-specialists. Mapping patterns were appropriate for the situation with STAR and STELLAR since there was a clear general use case – inter site cross search without requiring clients to possess extensive reasoning capabilities, with the focus on key archaeological concepts [22]. It is possible to define new patterns although this involves more technical expertise.

In some situations there may not be any clear use case that can be reflected in the patterns with which to drive the mapping. Sometimes the use case may emerge following more thorough reflection of the purpose of the mapping exercise. In other situations, it may be considered desirable to capture every aspect of the original dataset for unspecified and unknowable future research purposes. In this case, it may be harder to specify higher level mapping patterns but it should still be possible to specify lower level micro-patterns that can be combined together.

5 Future work

The recent specification by the CRM-Sig of definitive URIs for CRM entities has facilitated one aspect of implementation representation. We need to revise the STELLAR templates and the CRM-EH to conform to this.

We concluded our 2009 DAI workshop presentations with the following proposed issues to take forward, assuming they were considered possible and desirable:

- Agreement on implementation details (e.g. primitives)?
- Agreement on archaeological vocabulary approaches?
- Agreement on archaeological CRM extensions?
- Agreement on mapping patterns and guidelines?

In our view, these issues are still relevant today. We would also add additional aspects – the desirability of expressing the end-purpose of a mapping exercise; the provision of appropriate registries of mapping patterns; core metadata for mapping patterns together with the means for potential users to discover the patterns.

6 Acknowledgements

The STAR and STELLAR projects were supported by the Arts and Humanities Research Council [grant numbers AH/D001528/1, AH/H037357/1]. Thanks are due to the Archaeology Data Service for provision of the OASIS corpus and for many helpful discussions. Thanks are also due to Phil Carlisle (English Heritage), Andreas Vlachidis (University of South Wales) and the participants of the STAR and STELLAR workshops.

References

1. Patel, M., Koch, T., Doerr, M., Tsinaraki, C.: Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1. (2005)
2. CIDOC CRM: CIDOC Conceptual Reference Model. Heraklion, Crete: Institute of Computer Science, Foundation for Research and Technology. http://www.cidoc-crm.org/
3. Cripps, P., Greenhalgh, A., Fellows, D., May, K., Robinson, D.: Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper (2004), http://cidoc.ics.forth.gr/technical_papers.html
4. CRM-EH: English Heritage Extension to CRM for the archaeology domain, http://hypermedia.research.southwales.ac.uk/kos/CRM/ http://purl.org/crmeh
5. RDFS Encoding of the CIDOC CRM, http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs
6. STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.southwales.ac.uk/kos/star/
7. STELLAR Project. Semantic Technologies Enhancing Links and Linked data for Archaeological Resources. University of South Wales: Hypermedia Research Unit. http://hypermedia.research.southwales.ac.uk/kos/stellar/
8. Binding C., Tudhope D. 2009. Breaking Down Barriers to Interoperability. Interconnected Data Worlds: Workshop on the implementation of CIDOC-CRM, organised by the German Archaeological Institute in Berlin and funded by the TOPOI Excellence Project. http://www.dainst.org/medien/de/10_TudhopeBinding_STAR.pdf
9. Archaeology Data Service: Unpublished Fieldwork Reports (Grey Literature Library) http://archaeologydataservice.ac.uk/archives/view/greylit/
10. Binding, C., Tudhope, D., May, K.: Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus, 280–290. Lecture Notes in Computer Science, 5173, Berlin: Springer. (2008)

11. SKOS: Simple Knowledge Organization Systems - W3C Semantic Web Deployment Working Group `http://www.w3.org/2004/02/skos`
12. SENESCHAL Project: `http://hypermedia.research.southwales.ac.uk/kos/seneschal/`
13. Archaeology Data Service: Linked Data. `http://data.archaeologydataservice.ac.uk/`
14. CLAROS: The world of art on the semantic web. `http://www.clarosnet.org/`
15. CLAROS Wiki: CIDOC CRM entity description templates for Objects, Places, Periods & People `http://www.clarosnet.org/wiki/`
16. Nußbaumer, P., Haslhofer, B. (2007). Putting the CIDOC CRM into Practice – Experiences and Challenges. (Technical Report TR-200). University of Vienna. `http://cs.univie.ac.at/research/publications/publikation/infpub/404/`
17. Nußbaumer, P., Haslhofer, B., Klas W. (2010). Towards Model Implementation Guidelines for the CIDOC Conceptual Reference Model. Technical Report TR-201. University of Vienna. `http://eprints.cs.univie.ac.at/58/`
18. Dodds, L., Davis, I. (2012). Linked Data Patterns – A pattern catalogue for modelling, publishing and Consuming Linked Data. `http://patterns.dataincubator.org/book/`
19. Tudhope, D., Binding, C., May, K.: Semantic interoperability issues from a case study in archaeology. In: Stefanos Kollias & Jill Cousins (eds.), Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop (SIEDL) 2008, 88–99, associated with 5th European Semantic Web Conference, Tenerife (2008)
20. Binding, C., Tudhope, D.: Terminology Services. Knowledge Organization, 37(4), 287–298. (2010). Ergon-Verlag
21. Doerr, M. (2003). The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata, AI Magazine 24(3), 75–92
22. Tudhope, D., May, K., Binding, C., Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, Open access. `http://dx.doi.org/10.11141/ia.30.5`
23. Vlachidis, A., Tudhope, D. (2012). A pilot investigation of information extraction in the semantic annotation of archaeological reports. International Journal of Metadata, Semantics and Ontologies, 7(3), 222-235. Inderscience.
24. Tudhope, D., Binding, C., Blocks, D., Cunliffe, D. (2006). Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509-533. Emerald.
25. Tzompanaki, K., Doerr, M. (2012). A New Framework For Querying Semantic Networks. http://www.museumsandtheweb.com/mw2012/papers/a_new_framework_for_querying_semantic_networks Proc. Museums and the Web 2012.

Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories

Katerina Tzompanaki[1,2], Martin Doerr[1], Maria Theodoridou[1], Irini Fundulaki[1]

FORTH – Institute of Computer Science
N. Plastira 100 Vassilika Vouton
GR-700 13 Heraklion, Crete, Greece

[1]{katetzob, martin, maria, fundul}@ics.forth.gr
[2]{tzompana}@lri.fr

Abstract. Reasoning on provenance information and property propagation is of significant importance in e-science since it helps scientists manage derived metadata in order to understand the source of an object, reproduce results of processes and facilitate quality control of results and processes. In this paper we introduce a simple, yet powerful reasoning mechanism based on property propagation along the transitive part-of and derivation chains, in order to trace the provenance of an object and to carry useful inferences. We apply our reasoning in semantic repositories using the CIDOC-CRM conceptual schema and its extension CRMdig, which has been develop for representing the digital and empirical provenance of digital objects.

Keywords: Semantic networks, information access, semantic search, metadata, reasoning, provenance

1 Introduction

Over the last decade, semantic repositories that integrate heterogeneous data sources under semantic schemata such as *ontologies* have become an important component of the Semantic Web. These repositories usually support limited forms of *reasoning* that are used to infer *implicit knowledge* along subsumption relationships. Large-scale metadata repositories, i.e., semantic networks of RDF[1] triples integrating large amounts of data, have been developed and are globally accessible via the Internet. The list of such projects about cultural-historical data is long, including the Europeana[2], cultureSampo[3], German Digital Library[4], ResearchSpace[5], WISSKI[6], and

[1] http://www.w3.org/RDF/

[2] http://www.europeana.eu/

[3] http://www.kulttuurisampo.fi/?lang=en

[4] http://www.deutsche-digitale-bibliothek.de/

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

CLAROS[7] Projects. Linked Open Data[8] are advocated for cultural institutions, in which RDF data reside on local servers, and are accessible under published RDF schemata. In these systems, the CIDOC-CRM[9] [1] is becoming more and more popular as a rich RDF schema adequate to integrate complex cultural data.
These semantic repositories naturally follow the “Open World Assumption”, where knowledge is regarded as *incomplete* since metadata may be created by different people who state different facts about the same artifact and even may use the schema in different albeit correct ways. For instance, someone may say that an artifact is from Athens and someone else that the same artifact is part of the Parthenon Frieze in London. Thus establishing the correlation among information coming from multiple sources or even the same source becomes a necessity but simultaneously a great challenge. As in any Open World system, also in cultural heritage semantic repositories, users cannot know precisely what has been documented and how. So, while searching in the metadata they may ask for *implicit* knowledge like:

- characteristics (properties) of artifacts that have been recorded somewhere in the semantic network but are not directly associated to the object of interest [2]. For instance, the material from which an object is made of is recorded for the object parts and not for the object itself.
- characteristics that have multiple modeling alternatives. For instance, the “place of origin” of an object may be perceived as anything like its (a) place of creation, (b) place of discovery, (c) place of use and/or (d) creator’s birthplace
- characteristics that are generalizations of sets of more specific properties. For instance, the *“has met”* property [3-4] denotes the symmetric relation among items and people that were present in the same event, including time intervals and places. More specifically the *“has met”* property can be considered as the super-property of many properties, such as *“carried out by”* or *“used”* and their inverse ones.

In this paper we introduce a simple yet powerful reasoning mechanism based on inference and completion of metadata, as a means to help scientists query a semantic repository in order to trace and understand the source of their results, to reproduce results and to ease quality control of results and processes. Generalization and inferring of metadata from related objects is achieved by using the *propagation* of some object properties along the transitive part-of and derivation chains of information. We base our reasoning on a semantic repository which uses CIDOC-CRM[10] and its extension CRMdig[11] [5-6] appropriate for representing provenance. The implementation of this mechanism is feasible and indeed simplifies the querying process of scientists upon complex semantic repositories in the cultural heritage field and beyond [4]. The described framework has been applied in the framework of the European IP 3D-

[6] http://wiss-ki.eu

[7] http://explore.clarosnet.org

[8] http://linkeddata.org

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

[10] CIDOC CRM v5.0.4 Encoded in RDFS. http://www.cidoc-crm.org/rdfs/cidoc-crm

[11] CRMdig 3.0 Encoded in RDFS. http://www.ics.forth.gr/isl/CRMext/CRMdig.rdfs

COFORM[12], funded by the European Community (FP7/2007-2013, no 231809). In this project metadata describing the digital provenance for empirical 3D modeling and digitization processes are recorded along with metadata about the physical objects. Digital provenance data form deep chains of events connected by input-output, with up to tens of thousands of intermediary products that "inherit" many properties along the processing chains up to data about the digitized objects themselves. Using reasoning rules, we result in high recall rates, as not only explicitly documented properties but also derived properties across independently created metadata records can be combined for calculating the desired results, as long as referential integrity along these chains is preserved. In parallel, the Research Space project has also implemented this approach following our model.

This paper is organized as follows: we first review related work in Sec. 2 before introducing the reader to the problem in Sec. 3; Sec. 4 describes our approach; conclusions are provided in Sec. 5.

2 Related work

Data provenance is one kind of metadata that can be used to answer basic questions such as "*who created this artifact?*", "*where and when was this artifact created?*", "*when was this artifact modified and by whom?*" [7]. Provenance can support a large number of applications [8]: (a) *data quality & reliability*, (b) *audit trail* (c) *replication recipes* and (d) *attribution*. Provenance information can be used to determine the use of resources, to detect errors in data generation, that is to provide an *audit trail* for the data. Repeatability of experiments is an essential problem in scientific data management. Having fine grained provenance information about the processes used to create a data product, allows one to *replicate* the results of experiments in order to verify or debate scientific results. Knowing the author/creator of an artifact allows one to determine the ownership of data and hence liability in the case of errors (*attribution*) [9]. The problem of storing, accessing, and querying provenance has received a lot of attention in the last years. Research has focused in the areas of workflow and database systems which deal with different levels of provenance granularity regarding the type of data collected about a specific product (a data product or the result of a process).

1. ***Workflow systems***: A workflow can be a *process* (a series of steps that leads to the creation of a real world artifact) or a program (e.g., a series of computations that produce a data item). The provenance of a workflow *(coarse-grained provenance)* can be thought of as the entire history of the derivation of the result of the process [7], [10]. The information stored for the specific process can include the *different versions* of the *software* and the *hardware* used, the *agents* that were involved in the workflow chain (processes, human agents) and the *"things"* (e.g. data) employed by the processes. The ability to query the provenance of workflows allows users to explore and better understand results and enables knowledge re-use [7]. A large number of work-

[12] http://www.3d-coform.eu/

flow provenance models have been developed to represent provenance such as OPM [11], Provenir Ontology [12] and latest the W3C Recommendation Provenance Ontology (PROV-O) [13]. OPM and Provenir represent information of computational processes only, whereas PROV-O models provenance information that is generated by different systems and exchanged under different contexts.

2. ***Database systems***: At the other end of the spectrum, *data provenance (fine-grained provenance)* provides a detailed trace of how a piece of data has been obtained from a transformation process (i.e. query) [10]. Data provenance may indicate (a) the tuples involved in the computation of a result tuple (*why-provenance*) (b) where these tuples reside (*where-provenance*) (c) the *query operators* used to obtain the result tuple (*how provenance*) [14]. The above types of provenance have been extensively studied for relational databases and only recently for Linked Data [15].

Despite the research that has been conducted in the above topics there has been no explicit approach developed for representing and reasoning about provenance along the transitive part-of and derivation chains. The above approaches deal only with computational processes on digital artifacts whereas in our approach we are able to reason combining metadata of real world objects with metadata of digital objects and to deduct useful inferences with multiple applications such as maintenance of repositories of digitization products and completion of metadata by implicit knowledge, in applications where production chains comprise thousands of intermediates and dozens of final products without need to manage this redundancy in the repository explicitly.

3 The problem

It is quite common that a user might be interested in a property that is not explicitly documented for the object, but can only be implicitly inferred from related data. For instance, someone may search for things "*made from: steel*", when objects may have been registered as *having parts* (using the "*is composed of*" property) that are "*made from: steel*". From this part-of property chain, we can deduct that the "whole" object is also made from steel, because it has parts made from steel. Moreover, the information may be represented in a different way than the one the user expects, for example instead of "*made from: steel*", objects may be defined with "*has type: steel object*". As the making of CIDOC-CRM demonstrated, it is impossible to normalize a global model for information integration to one unique representation for each property. Rather, in aggregation systems and the Semantic Web, one has to accept that properties are represented by sets of reasonable alternatives that can be related to each other by deductions.

The more analytical and precise a global model is, the less obvious it is for the user how a simple, intuitive question relates to the ontology. Transitive properties (such as parts of parts or derivatives of derivatives) cause "propagation" [16] of properties along those property paths. Propagation may be very complex to formulate as query, but is also very powerful when it comes to query recall improvement. For instance, one could assume that the actors, place and time that are reported for the building of

Parthenon (the “super-event”) also apply for or include the building of its friezes (a “sub-event”); or that materials a frieze is made of, are considered to be among the materials the whole Parthenon is made of; or that the subjects a frieze represents also apply to its copies or derivatives, etc. Such reasoning allows for inferring facts that are not stated within a single metadata record. Take for example the following information taken from two different sites. On one hand, we have the British Museum[13] website saying that the object with the description “Horsemen from the west frieze of the Parthenon” is part of the Parthenon, and on the other hand, there is the Acropolis Museum[14] stating that Parthenon was created by Pheidias. Using the CIDOC-CRM schema (prefixed with “crm”) the metadata describing these pieces of information are:

- “Horsemen from the west frieze of the Parthenon” *crm: forms part of* “Parthenon”
- “Parthenon” *crm: was produced by* “Construction of Parthenon” *crm: carried out by* “Pheidias”

Using reasoning on the integrated metadata we could infer that Pheidias was involved in the making of the Horsemen as well. In other words, in a query about the maker of the Horsemen, Pheidias would be deducted as a plausible answer. Thus, flat queries that do not take into account such inferences are more likely to have poor or even empty results. In another perspective, metadata built without including such inference rules, provide poorer knowledge. Such inference takes advantage of the transitivity property of *crm: forms part of* and *crm: carried out by* [2] and combined with application dependent relevance criteria can improve significantly the query results in specific application domains.

In provenance data, property propagation along part-of hierarchies can be observed between complex processes and their individual actions, between measurement devices and their components, between digital products and their parts. It must clearly be understood that virtually **none** of these inferences holds in a strictly logical sense. There is a **likelihood** for instance that the same lense of my camera was used throughout an image capture if not stated otherwise. Therefore all inferences we describe increase **recall** with respect to the documented reality**,** even though the mechanism is not an information retrieval technique. Assessing the respective probabilities is not the target of this paper and may be due to future work.

In the next section we propose a framework that utilizes rules to derive useful deductions about transitive properties, based on property propagation in cultural heritage semantic networks.

4 Reasoning using provenance information

Up to this point, we have discussed the necessity of a mechanism to reason upon complex structured metadata. In this section we propose such a mechanism that takes

[13] http://www.britishmuseum.org/

[14] http://www.theacropolismuseum.gr/en

advantage of the property propagation along transitive derivation and part-of chains, in order to derive useful inferences. Our priority is to improve query recall and resolve relevance issues with additional application specific constraints. To help the user understand the meaning and practical usefulness of the framework, we present it in the context of exploiting semantic networks and completing metadata. For this reason, we also include a set of real research questions from the Cultural Heritage domain that have been analyzed in terms of queries in the 3D-COFORM project metadata repository that consists of a semantic network containing rich cultural information [17] and supports the study of such research topics. Here we show that they can be answered easily with semantic, associative queries that make use of the proposed rules. The 3D-COFORM metadata repository consists of 1M RDF triples and is the result of over one year of intensive work, testing and validating the semantic reliability regarding the inference results of our conceptual modeling. We used the BigOWLIM reasoner and query optimization was achieved by implementing shortcuts for certain paths and defining specific reasoning rules. The proposed approach is also studied and validated in fields such as geology and biology.

Assuming that the reader is familiar with the basic semantic web notions, we attach to each query its graphical representation using terms from the CIDOC-CRM that adopts the following notation: Boxes represent classes, the upper part of which is the name of the CIDOC-CRM class (orange) or CRMdig class (blue) and the lower part is the value of an instance of that class, either fixed or represented by a variable. Arrows connecting two boxes denote properties between the two respective classes, and the name of the property is printed over the arrow. Variables are represented with the letters X, Y, Z, U, V, W and denote any node of the metadata graph fitting the respective path. Query parameters include terms, numbers, dates, and strings. The variables that are returned by the query are denoted with variables prefixed with '$', e.g. *$Material*, *$Monument, $Height*. We are now ready to introduce the first rule, which is based on the transitivity of properties in *part-of* chains.

Rule 1: *The property of an object is the aggregation of the explicitly defined property in the object itself and the respective properties of all its subparts.*

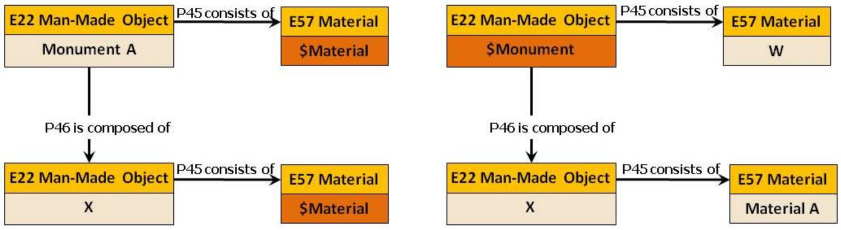


Fig. 1. Forward and backward traversal of the *part-of* chain

According to Rule 1, we can do reasoning by traversing the *part-of* chain either forward or backward (Fig. 1) and we can answer queries such as:

1. Find the material of Monument A: The material of Monument A includes its explicitly stated materials but also the materials of its parts. The query will forward traverse the *part-of* chain and collect all the Materials that have been registered both to Monument A and its parts.

2. Find Monuments constructed from Material A: The information regarding the material of an object might be registered in its parts and not directly in the object itself. So the query should search both the explicitly stated materials of the object and the materials of its parts too.

Fig. 2 presents an example of a monument which is composed of four subparts made of different materials. The object (statue of Queen Victoria[15]) does not have material information in its immediate, explicitly defined metadata but its subparts do have. Our reasoning approach will include this object in the answer set of the query "*Find all statues made of Bronze*" whereas queries relying only on explicitly defined metadata, would fail to retrieve it. Similarly, with our approach, the answer set of the query "*Find the material of the Queen Victoria Monument*" is {Grey granite, Grey marble, Bronze} while the traditional query would get an empty answer set. Using property propagation results in high recall rates however a statistical factor that may deteriorate precision is introduced, since a property is not necessarily propagated along a path or it's significance is not important. For example consider the case of The Kissing Bridge[16] sculpture, which is composed of, (i) two bases made of concrete, and (ii) two statues made of bronze. The significant information in this case is that the statues are made of bronze. Our reasoning approach will influence recall since we will infer that the Kissing Bridge sculpture is made of concrete and bronze. Precision can be improved by adding constraints on the queries.

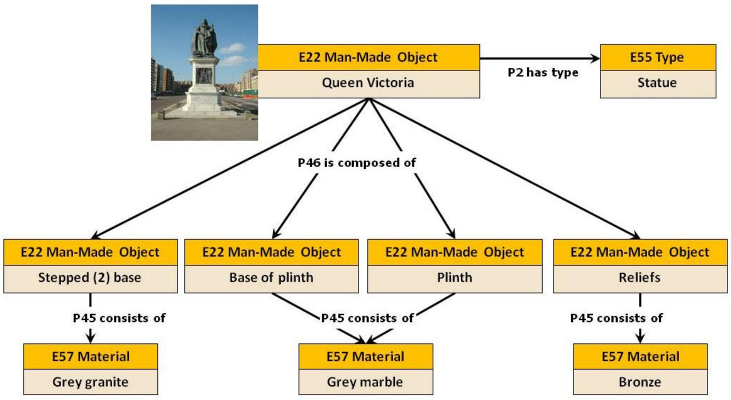


Fig. 2. An example of Queen Victoria statue metadata

[15] Public sculptures of Sussex http://www.publicsculpturesofsussex.co.uk/object?id=71
[16] Public sculptures of Sussex http://www.publicsculpturesofsussex.co.uk/object?id=127

Except from the part-of chains, the derivation chains can also be used for transfer of properties among material and immaterial objects. More specifically, CRMdig Digitization Process class marks property transfers from physical to digital objects while CRMdig Formal Derivation class marks property transfers from digital to digital objects [6]. We make the assumption, that the transformation of a physical object to its digital representation is achieved through "subject preserving" events, which means that the physical object depicted in the derivatives remains the same as the one in the derivation source. Based on this principle, we proceed to our second rule below.

Rule 2: *Physical objects may share properties with their digital representations and their derivatives.*

According to Rule 2, we can do reasoning by traversing the derivation chain (Fig. 3) either forward or backward and we can answer queries such as:

1. Find objects that depict Actor A: Physical Object A has an explicit declaration of the depicted Actor A. This property is propagated to the digital representations of Object A and thus we can infer that all Data Objects (X, Y, … Z) depict Actor A.

2. Find the size of Object A: An object's 3D model may have the size of the object automatically calculated and stored in its metadata. This property is backwards propagated through the derivation chain and thus we can infer the size of the physical object through the size registered in the metadata of its 3D representation.

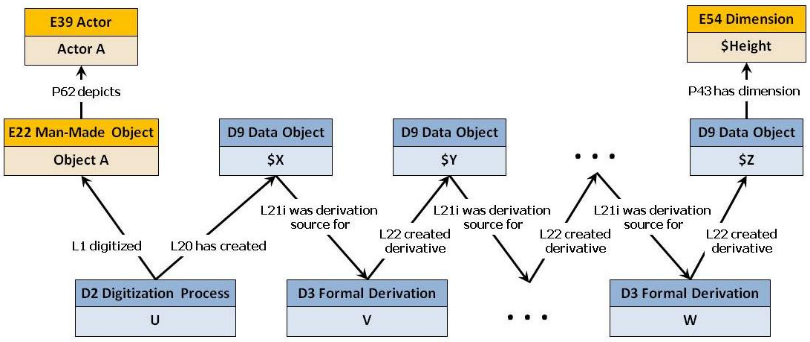


Fig. 3. Property propagation along the *derivation* chain

Fig. **4.** presents an example of a statue that depicts Ramesses II. The statue has been laser scanned and processed by MeshLab to produce its 3D model. The object "Ramesses Statue 1" does not have any size information in its immediate, explicitly defined metadata. However, our reasoning approach can answer the query "*Find the size of the Ramesses Statue 1 Object*" by retrieving the size calculated in the "3D model of Ramesses Statue 1" object and inferring that it also applies to the original physical object. Similarly, with our approach, the answer set of the query "*Find all the objects that depict Ramesses II*" is {"Ramesses Statue 1", "Scanned Ramesses

Statue 1", "3D model of Ramesses Statue 1"} while a query without inference capabilities would retrieve only {"Ramesses Statue 1"}.

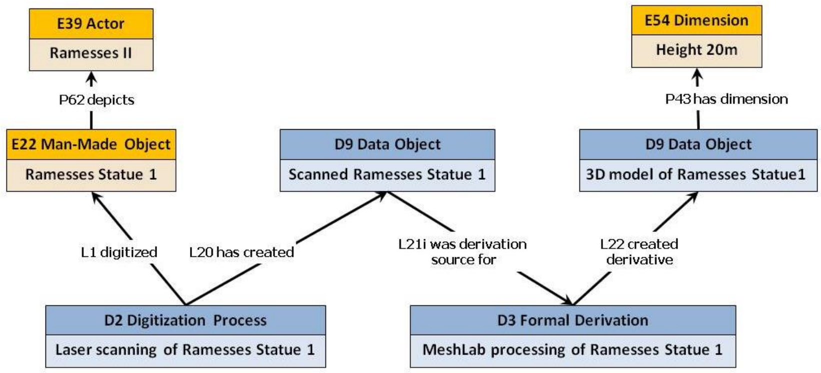


Fig. 4. An example of property propagation along the *derivation* chain

The combination of the property propagation along the two chains described above can help solve research questions that cannot be answered without reasoning. Consider the following research question: "Find Temples where Ramesses II and his wife Nefertari have the same size". If we apply both our rules on the metadata graph displayed in Fig. 5, we will get the set {"Abu Simbel Temples", "The Small Temple"} as an answer to our research question.

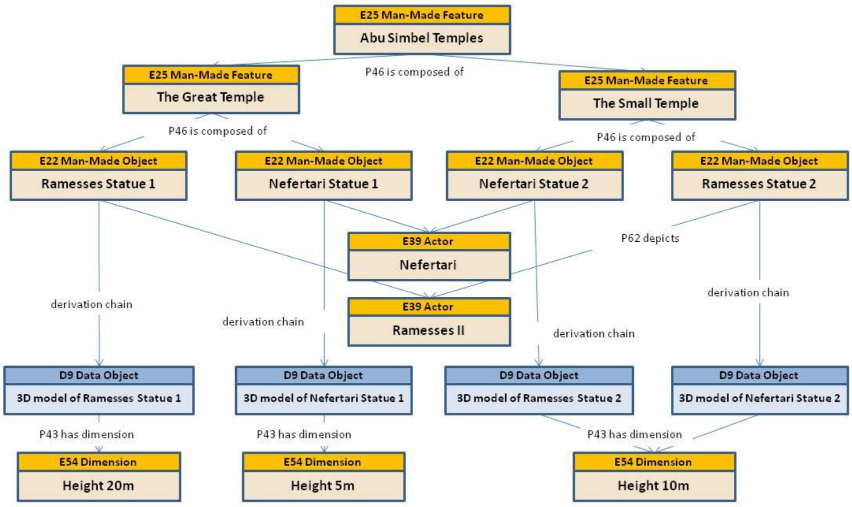


Fig. 5. Property propagation along the derivation and part-of chains

Here we have displayed two basic rules that can be used in a variety of applications, like quality control or querying. We encourage readers especially interested in the application of reasoning rules for querying purposes, to refer to the technical report in

[18] for a thorough study of the matter. As reported in [18], an exhaustive set of such rules has been implemented and tested by our team. The number of necessary rules is considerably reduced by property subsumption, but nevertheless we had to produce over a hundred counting all combinations.

5 Conclusion

In this paper, we have demonstrated a simple yet powerful mechanism of reasoning on provenance information by propagation properties along derivation and part-of chains. Moreover, we report an implementation on metadata built on the CIDOC-CRM and CRMdig schemas in the cultural heritage domain. In this implementation, it can be verified that the combination of structuring the metadata with rich schemas and applying reasoning upon them leads to the deduction of useful inferences with multiple usages. A number of such example use cases can be listed: (1) maintenance of repositories of digitization products, (2) garbage collection on reproducible intermediate files, (3) trace dependencies of products on tools and algorithms that should not become obsolete for long time preservation, (4) (re)production of valid, complete metadata at a loss of intermediate files, (5) completion of metadata by implicit knowledge, when production chains comprise thousands of intermediates and dozens of final products without need to manage this redundancy in the repository explicitly.

6 References

1. Doerr M.: The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata, AI Magazine, Volume 24 (3), pp.75-92 (2003)
2. Strubulis, Ch., et al.: Evolution of Workflow Provenance Information in the Presence of Custom Inference Rules. *SWPM2012-Proceedings of the 3rd International Workshop on Semantic Web in Provenance Management.* May 28, Heraklion, Greece (2012)
3. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini C., van de Sompel H.: The Europeana Data Model (EDM). *World Library and Information Congress: 76th IFLA General Conference and Assembly.* August 10-15, Gothenburg, Sweden (2010)
4. Tzompanaki, K., & Doerr, M.: A New Framework For Querying Semantic Networks. *Museums and the Web 2012,* San Diego, CA, USA (2012)
5. Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., Melessanakis, V.: Modeling & Querying Provenance by Extending CIDOC CRM. Distributed and Parallel Databases, Vol. 27 (2), (2010)
6. Doerr, M., & Theodoridou, M.: CRMdig: A generic digital provenance model for scientific observation. *TaPP'11, 3rd USENIX Workshop*. June 20-21, Heraklion, Crete (2011)
7. Davidson, B. and Freire, J.: Provenance and Scientific Workflows: Challenges and Opportunities. *SIGMOD*, (2008) (Tutorial Track).
8. Goble, C.: "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics", Workshop on Data Derivation and Provenance, (2002)
9. Simmhan, Y. L. and Plale, B. and Gannon, D.: A Survey of Data Provenance in e-Science. *SIGMOD Record* Vol. 34 (3), (2005)
10. Tan, W-C.: Provenance in Databases: Past, Current and Future. IEEE Data Eng. Bulletin 30(4), (2007)

11. Moreau, L. et al.: The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, 27(6), (2011)
12. Sahoo, S., Thomas, C., Sheth, A., York, W.S., and Tartir., S.: Knowledge modeling and its application in life sciences: a tale of two ontologies. In Proceedings of WWW, (2006)
13. Lebo, T., Sahoo, S., McGuiness, D.: PROV-O: The PROV Ontology. W3C Recommendation, (2013). Available at http://www.w3.org/TR/prov-o/.
14. Cheney, J., Chiticariu, L., and Tan, W.-C.: *Provenance in Databases: Why, How, and Where.* Foundations and Trends in Databases. Vol 1(4), (2007)
15. Theoharis, Y., Fundulaki, I., Karvounarakis, G., and Christophides, V.: *On Provenance of Queries on Semantic Web Data.* IEEE Internet Computing 15(1), 31-39 (2011)
16. Wickett, K.M., Renear, A.H., Urban, R.: "Rule Categories for Collection/Item Metadata Relationships" In *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology.* Pittsburgh (2010)
17. Doerr, M., Tzompanaki, K., Theodoridou, M., Georgis, Ch., Axaridou, A., & Havemann, S.: A Repository for 3D Model Production and Interpretation in Culture and Beyond. *Proceedings of VAST 2010: The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*. 21-24 September, Paris, France (2010)
18. Tzompanaki, K., & Doerr, M.: FORTH-ICS Technical Report TR429, Fundamental Categories and Relationships for Intuitive querying CIDOC-CRM based repositories (2012)

Representation of Archival User Needs using CIDOC CRM

Steffen Hennicke[1]

Berlin School of Library and Information Science,
Humboldt-Universität zu Berlin, Germany
steffen.hennicke@ibi.hu-berlin.de

Abstract. This paper stems from an ongoing dissertation project and demonstrates how the CIDOC CRM is used to create an ontological model – the *Archival Knowledge Model* (AKM) – of common patterns found in written natural language questions to archives. Such an ontological model can be used to query archival or historical knowledge bases in order to provide more adequate answers and to enable more relevant discovery facilities. For this purpose, 330 reference questions to the German Federal Archive are being analyzed and patterns found translated to the CIDOC CRM and appropriate extensions. In particular, the paper introduces the methodological approach to the interpretation of user questions and the draft of a prominent pattern called *Documentation-Activity.*

Keywords: CIDOC CRM, archival reference questions, access to archives, archival user needs, Archival Knowledge Model

1 Introduction

The main means for discovering [1] and accessing primary sources in an archive are finding aids and holding guides supported by the expertise of archivists. These archival aids are descriptive tools which are meant to help the user to locate and discover relevant materials in the enormous and ever growing amounts of rich *information potentials* [2] in archives. The conceptualization of these descriptive tools as well as respective digital encoding standards like the *Encoded Archival Description*[1] (EAD) are based on elaborated and historically grown archival principles and models but their design is less informed by explicit knowledge about the information needs of archival users [3] due to a prevailing lack of qualitative in-depth analysis of archival user needs [4][5]. Such studies are, however, a crucial cornerstone for the improvement of existing and future digital archival information systems [6].

The hypothesis on which this paper rests is that it is not necessary, and even not desirable, to change the archival description itself and related metadata schemas, but, instead, it is possible, in principle, to supplement existing archival

[1] http://www.loc.gov/ead/

and historical knowledge bases with an ontological model which matches typical patterns from user inquiries to archives. Such an ontological model would make knowledge explicit and add relevant context which is necessary to adequately answer typical user questions and to create better discovery systems. Furthermore, such ontological models enable empirically qualified assessments of metadata schemas for archival information systems but also of archival cataloging rules.

The general research question, therefore, is if there is a hypothetical ontology which can represent user inquiries and their probable interpretations as formal queries against a model of the archival target world that would adequately answer the inquiry or its implicit purpose. The result of this analysis is an ontological model which represents inquiry patterns of different abstraction levels to archives in the form of queries to this ontology. The CIDOC CRM[2] has been chosen as the ontological target model mainly for its strong empirical foundation and event-based conceptualization of historical processes. Written *reference questions*[3] from the German Federal Archives, the *Bundesarchiv*, have been chosen as research data. This type of research data has been largely neglected in the analysis of user needs in the archival domain, although they document a mostly unfiltered information need in the users own words.

A brief literature review will establish the general research context followed by a short introduction of the research data and the methodological approach to the analysis of questions.[4] The focus of this paper lies on the demonstration of the interpretative translation of natural language questions to a common ontological representation. Two examples will demonstrate how shared patterns in user questions and their probable interpretations can be translated to an ontological model, the *Archival Knowledge Model* (AKM), covering and extending the CIDOC CRM. The specific pattern presented in this paper is called *Documentation-Activity* which proposes two new classes as extensions to the CIDOC CRM. It is important to bear in mind that all results presented in this paper are preliminary and research is ongoing.

2 Research Context

The limitation on "simple answers to clear cut, search term-based questions" [7] is one of the core problems of today's information systems on the Web. Pattern-oriented retrieval could describe many more complex questions whose answers go beyond the capacity of simple querying [8]. This limitation poses a significant

[2] http://cidoc-crm.org/

[3] The term reference question refers to a request of a user to a staff member of a library or archive for information or assistance regarding the provision of any kind of information. Such a request can either be posed in person at a reference desk or remotely by phone, mail, or e-mail. In this study, only written reference questions by mail or e-mail are being analyzed.

[4] For more details on the dissertation, please confer the extended abstract which will be presented at the Doctoral Consortium of the TPDL 2013 and published in the conference proceedings.

barrier to more sophisticated and integrated information systems. Part of this problem is a prevalent focus on traditional library cataloging and methodology in describing and contextualizing objects of interest. At the same time, today "the key challenge of organizing information is to construct systems that aid understanding, contextualizing, and orienting oneself within a mass of resources" where models help to bridge a semantic gap between the formalizations in information systems and the conceptualizations of scholars [9]. Instead of a *Global Knowledge Network* [7], mostly isolated "silos of information" exist which all employ their own idiosyncratic structures and data encodings. The Semantic Web addresses these issues in its research agenda. However, this agenda suffers from an "almost exclusive focus on 'terminology' rather than 'ontological structures'" resulting in the neglect of fundamental and complementary lines of research [7]. One such missing line of research is how typical user questions are formally structured. The systematic and in-depth analysis of original user questions from different stages of the research process is important and has the potential to provide, for example, necessary information on query mechanisms or adequate granularity of ontologies [7].

Discovery is one of the most important and re-occurring stages in research processes especially distinctive for historical inquiry in archival settings. As already mentioned, research in the archival domain exhibits a lack of in-depth user studies [10]. The study of Duff and Johnson [11] is one of the few which looked at the type and structure of user reference questions to archives.[5] Regarding the domain of historical research, Case [15] concluded that history "may be less well served by classification and indexing than any other academic field" and that the "accomplished scholar - and particularly the historian - is not often aided by the disciplinary boundaries that library classification schemes enforce." Instead of the "disciplinary model of a body of knowledge, subdivided by place and period", the so called "problem-oriented model" should be used as the basis for the design of future tools and services for historians. At the same time, Case correctly points out that it is not viable to fundamentally change documentation practices and reorder collections of archives and libraries but that special services and tools might be able to bridge (semantic) gaps between the user and existing knowledge bases.

3 Research Data

The *Bundesarchiv* is the Federal Archives of Germany who are responsible for the permanent preservation and accessibility of federal archival documents from the civil and military archives of the Federal Republic of Germany (since 1949) and its predecessors. In addition, significant documents of private origins and from political parties, associations and societies are kept in the archive. The number of written inquiries to the Bundesarchiv amounts to roughly 60,000 per

[5] Similar studies are, for example, from Collins [12], Conway [13], and Gagnon-Arguin [14].

year, based on the numbers from 2008 and 2009.[6] The Bundesarchiv has granted supervised access to their user files which contain a physical documentation of the correspondence and interaction between a user and the archive. Each user file carries an identification number which is retrievable through a database system offering a small range of search facets[7] related to the user and associated user files. Based on these facets a sample of 196 user files was retrieved, which was further complemented by a special selection of 40 rich user files which had been collected by the head of the department *Stiftung Archiv der Parteien und Massenorganisationen der DDR im Bundesarchiv* (StA). The sample of user files shares as a general historical and topical horizon Contemporary German History (19th and 20th century) and contains rich and challenging inquiries. The sampling process was informed by educated assumptions, professional advice of archivists, and skimming through user files. The collection was stopped when the questions extracted from the user files appeared to exhibit no new qualities or substantial variances. Reliable information about the users' background was not available.

In total, 236 user files have been selected from which 100 were available for further study. Only 60 user files contained at least one explicit or implicit information request as part of an inquiry by email or letter. From these 60 user files, 546 single questions were extracted and pre-analyzed[8] according to the methodology of Duff and Johnson [11] with very similar results. Regarding the type of question, 260 questions were of type "explicit" or "implicit resource discovery" (material-finding, specific form, specific item, consultation), 70 questions were "factual", and 216 questions consisted of "administrative/directional", "user education", or "service request" questions. The questions of the type "resource discovery" and "factual", in total 330 questions, are part of the discovery stage in the research process and are currently being analyzed as described in the following sections.

4 Methodological Approach

The methodological approach taken in this study goes beyond the analysis of the mere utterance level and syntactic structure of the inquiry and focuses on the interpretation of the questions. Here, the sense of an inquiry is interpreted in order to discover the implicit questions with regard to a certain domain of discourse. In the scope of this work, two domains of discourse are being distinguished: the *archival domain* of record keeping and the *domain of historical*

[6] http://www.bundesarchiv.de/oeffentlichkeitsarbeit/publikationen/taetigkeitsberichte/

[7] This includes, for example, the general purpose of the inquiry as given by the user on the user management form, a general subject and time frame of the inquiry's topic, or the department initially responsible for processing the inquiry which allows concluding on the origins of the archival material. However, it is important to note that these classifications are coarse and not meant for precise retrieval of user files based on these search facets.

[8] The publication of the results is in preparation.

inquiry for which traces and evidence can be expected to be found in the archive. These two domains constitute the epistemological baseline for the interpretation of the inquiries: What might the user need to know in order to satisfy his research interest? Reality is then described in a way so that it fits the perceived epistemological interest of the user and his question. This process is necessarily an act of interpretation and relies on educated intuition regarding both domains and necessarily filters probable implicit questions. It does not, to be sure, aim at "truthful" models in terms of some perceived "objective" meaning or structure of a question. Regarding such epistemological issues of interpretation in relation to historical science and theory of history, the approach to interpretation taken here understands itself as *meta-theoretical*, similar to Gardin [16] in the domain of archeology. It is agnostic to specific types of historical sciences but reflects patterns applicable to general historical inquiry.

The patterns which are identified in the questions are modeled in CIDOC CRM which describes historical facts in terms of possible relations between universals. It is the result of an empirical analysis of existing conceptualizations of the cultural-historical world in the form of metadata structures. One of the most important design principles of the CIDOC CRM is to represent the past as discrete events. Material and immaterial persistent items are present at events either as a concept or via a physical information carrier. History, therefore, is conceptualized as meetings of persistent items through events in space-times [17]. The historical-archival domain of the analyzed inquiries is in the scope of the CIDOC CRM. For these reasons, its methodology is adopted in this work and it is tried to identify whether the CIDOC CRM will completely or partially cover the hypothetical ontology. In the latter case, appropriate extensions to the CIDOC CRM will be proposed.

Formally, an ontology engineering approach is employed in that the inquiries and their interpretations are being translated to an ontological model based on the CIDOC CRM and appropriate extensions.

5 Translating Patterns of Questions to CIDOC CRM

Two examples will motivate how questions are being analyzed and how their interpretation is formally represented in an ontological model based on the CIDOC CRM. An inquiry typically consist of contextual information and one or more direct or indirect questions.[9]

5.1 Example 1

Context: *One source I would like to consult are the police- and surveillance reports for the Weimar Republic which are about revolutionary movements. I would*

[9] All questions have been translated from German to English by the author of the paper. Text in square brackets has been added either to make named entities anonymous or to clarify the meaning of certain paragraphs. Finally, red borderlines indicate the entity at which a question is targeted.

like to know what the surveillance agency of the Reich (or the ones of the Länder) had to say about [person name].

Question 1: *Do you know if the Bundesarchiv holds such documents?*

Question 2: *Which agency of the Reich was responsible for the surveillance of the revolutionary movements? The Reich or the Länder?*

The *given* elements in these two questions and their context are the name of a specific actor ("[person name]"), the type or function of a group ("revolutionary movements"), the type or function of a legal body ("surveillance agency of the Reich"), the type or function of documents ("police- and surveillance reports"), and the name of a period ("Weimar Republic").

The interpretation of the questions can be structured into two principle steps. The first one is concerned with the *wanted* information asking for the research interest of the user's question: Which are probable or adequate answers to the question with regard to the domain of historical inquiry but also to the archival domain?

In the case of the first question the user is looking for reports which are the result of a policing or surveillance activity targeted at a specific type of group ("revolutionary movements") or at a specific person ("[person name]"). In that way, the first question could be even seen as a two-fold question. The results of these policing or surveillance activities are documents about the activities of the aforementioned actors. Such documents as routinely products of a governmental institution are now stored in an archive. The user wants to know if such documents are available in the Bundesarchiv. Therefore, the information the user wants are pointers to appropriate documents, for example, call numbers of files likely to contain relevant documents.

The second question in the example is a fact-finding question. It operates with the same given information but asks for a different wanted information. The user wants to know which agency was generally responsible for surveillance activities targeted at a specific type of actor. He is inquiring for a name of one or more legal bodies. The word "responsible" is important because it stresses the fact that whatever agency conducted the surveillance activities did so following a mandate which formally delegated said responsibility to the agency.

The second interpretation step comprises the translation of the question, its context and its interpretation to the CIDOC CRM. The CIDOC CRM suggests that historical facts and entities are related to each other through events which form the world lines in history. Therefore, the second interpretation step asks how the given and wanted information entities relate to each other.

The first two-fold question can be represented in CIDOC CRM as shown in figure 1. This is a simplified representation expressing the formal basic structure of an answer adequate to satisfy the wanted information or the research interest.[10] The interpretation of the question is evident and materialized by the

[10] The implicit question for pointers to documents, for example, a set of call numbers, is not the point when translating to CIDOC CRM but the *context* of the documents of interest. Identification for retrieving the actual physical document is not in the scope of this ontological model.

documentation activity in the center of the figure. This class is a proposed extension to the CIDOC CRM and will be introduced in more detail later on. The documentation activity is seen as being implicit in the historical reality referred to in the question: The police- and surveillance reports have been created during an event, or a series of events, which ”documented” some events which are qualified by the participation of an actor (”[person name]”) or a specific type of group (”revolutionary movements”). The documentation activity is following a mandate which captures a specific type of ”documented plans (...) for deliberate human activities”.

Most importantly, mandates specify or govern documentation activities. This class is another proposed extension to the CIDOC CRM and will also be introduced in more detail later on. In the case of the first two-fold question the mandate has a specific type of group as its principle target and at the same time aims at a specific actor. Furthermore, the mandate is assigned to an actor, in this case an institution, who carries out the actual documentation activity which, as the last relevant contextual information, falls within in the historical period of the Weimar Republic. Documents which are the result of this constellation are relevant documents and may adequately answer the user’s first two-fold question.

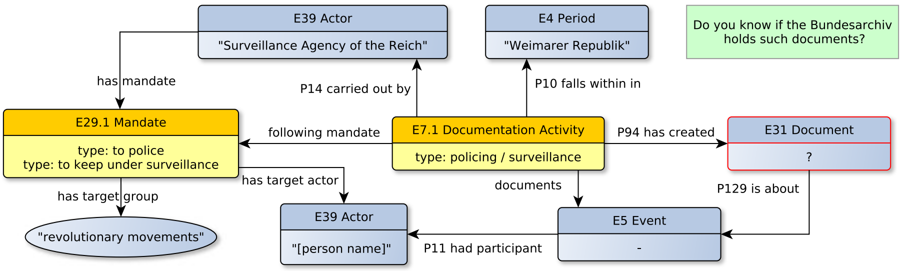


Fig. 1. Question 1 in AKM

Figure 2 shows the translation of the interpretation of the second question to the CIDOC CRM. An adequate answer can be modeled within the same pattern as for the first question. In this case the wanted information is the name of an actor who had the mandate to police or to keep under surveillance revolutionary movements during the Weimar Republic.

These two questions and their representations in CIDOC CRM show a common core pattern which is grouped around a documentation activity which documents events and which is following a specific mandate. This relation between documentation activity and mandate is essential. It can be identified in many other questions through interpretation.

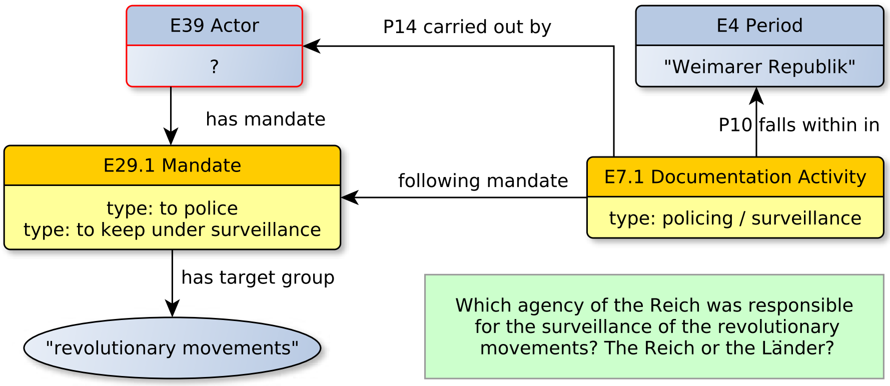

E39 Actor
?
P14 carried out by
E4 Period
"Weimarer Republik"
has mandate
P10 falls within in
E29.1 Mandate
type: to police
type: to keep under surveillance
following mandate
E7.1 Documentation Activity
type: policing / surveillance
has target group
"revolutionary movements"
Which agency of the Reich was responsible for the surveillance of the revolutionary movements? The Reich or the Länder?

Fig. 2. Question 1 in AKM

5.2 Example 2

The second example shows that seemingly different questions exhibit very similar patterns and that documentation activities based on mandates may cover a broad range of different types of activities. Furthermore, some finer notions like self-documentation and documentation of others are introduced in this example.

Context: *In 1980, a delegation of the FDGB lead by Harry Tisch laid down a wreath of flowers in Oradour. The visit was part of a trip of the FDGB to France (demonstration in Limoges, reception and meeting with the FKP and CGT in Paris). At this time, Tisch was also a member of the Politbüro of the ZK of the SED.*

Question 1: *Where can documents be found about the planning [of this trip]...*

Question 2: *...and the report on this trip?*

Question 3: *In your opinion, has such a trip been discussed or, at least, been approved in the ZK?*

The first question asks for documents about the planning of the trip to France while the second question asks for the report on this trip. In both cases the documents refer to the same event "Trip to France" but they are the result of two distinct activities. The first one, the planning activity, happens prior to the actual trip and does not directly document the trip but series of planning events. The second documentation activity, the reporting activity, produces one or more documents which report on the trip event itself. Both questions ask for pointers to documents as the result of their respective documentation activity.

Figure 3 combines the first and second question and their interpretations. The documents are the result of documentation activities which document events which were involved in the planning of the trip to France. In the case of the second question, the documents are the result of a documentation activity reporting on the event "Trip to France". Necessarily, both documentation activities followed a mandate to do so and were carried about some actor.

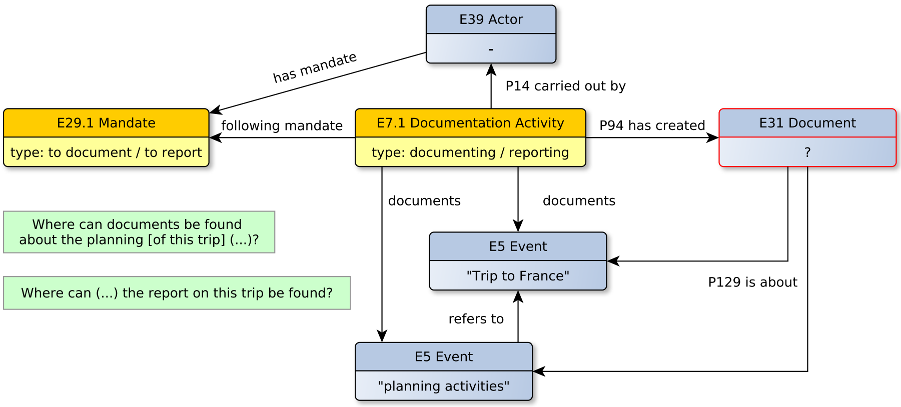

E39 Actor
-
has mandate
P14 carried out by
E29.1 Mandate
type: to document / to report
following mandate
E7.1 Documentation Activity
type: documenting / reporting
P94 has created
E31 Document
?
documents
documents
Where can documents be found about the planning [of this trip] (...)?
E5 Event
"Trip to France"
P129 is about
Where can (...) the report on this trip be found?
refers to
E5 Event
"planning activities"

Fig. 3. Questions 1 and 2 in AKM

The third question should adumbrate some more difficult issue in terms of interpretation and translation of questions. The question asks if a specific actor, the "ZK"[11], had discussed or approved a specific event, the trip to France.

First of all, it is important to remember that the patterns are about the general and generic relations between certain entities and not about the many specific qualities of these connections: it is not relevant if the relation between a document and an event is one of "discusses" or "approves" but that, on the most generic semantic level, it is a relation of "aboutness". It is the genuine task of the researcher to read and interpret the documents in order to find out about the qualitative aspect if the ZK did in fact "discuss" and, even more, did "approve" something. The pattern is a means for the researcher to discover potentially relevant documents. One tentative inference which might be drawn from a knowledge base which instantiates this ontological model is that the ZK, or more precisely some members of this group, must have had knowledge of the event "Trip to France".

Therefore, relevant documents for an adequate answer include those ones which are about events during which the actor ZK was present and which in some way refer to the event "Trip to France". An example for such an event could be the planning event from the first question. Figure 4 shows another possible scenario where the ZK carried out a committee meeting during which the trip to France has been mentioned and which has been documented through minutes.

Again, the minutes are the result of a documentation activity which follows a mandate to take minutes. In this case, the ZK is the actor who not only follows this mandate and carries out the documentation but also conducted the event which is being documented. This is a kind of *self-documentation* as opposed

[11] "ZK" is the abbreviation for *Zentralkomittee* ("central committee") which belonged to the *Sozialistische Einheitspartei Deutschland* (SED).

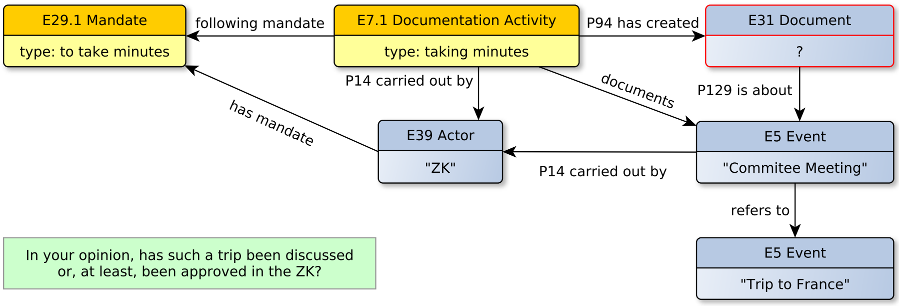


Fig. 4. Questions 3 in AKM

to the documentation of others in the case of surveillance and will be briefly discussed in the next section.

6 The Documentation-Activity Pattern

The examples previously discussed exhibit a shared pattern which is able to accommodate a broad range of different questions and their probable interpretation in terms of adequate answers. This section will introduce the current draft of the *Documentation-Activity* pattern as shown in figure 5. So far, this pattern appears to be one of the most prominent and complex results from the analysis of the user questions.[12]

At the core of this pattern resides a new proposed class *E7.1 Documentation Activity*. This new event class is an extension to the CIDOC CRM in order to appropriately capture the essentials of activities which, literally speaking, document *E5 Events* and which create one or more *E31 Documents*. It is a sub-class of *E65 Creation* and not of *E7 Activity* because a characteristic feature of the documentation activity is the creation of documents and only events of the type *E65 Creation* "result in the creation of conceptual items or immaterial products" through *P94 has created*. Furthermore, the scope of *E7.1 Documentation Activity* is more specific than that of *E65 Creation* in that documentation activities document *E5 Events* and, most importantly, follow a mandate. The representation of the fact that a documentation activity follows a mandate led to the introduction of a new property called *follows mandate*.

The *E29.1 Mandate* is the second proposed extension to the CIDOC CRM as a sub-class of *E29 Design or Procedure*. The mandate formulates the principle scope of application for documentation activities in that it specifies who has the mandate to execute the documentation activity and which specific actors, types

[12] While the analysis of the questions is on-going and no reliable evidence based on the current research sample can be provided at this point, an estimate of at least 30% of all questions in the sample might be adequately covered by this pattern either partially or in full.

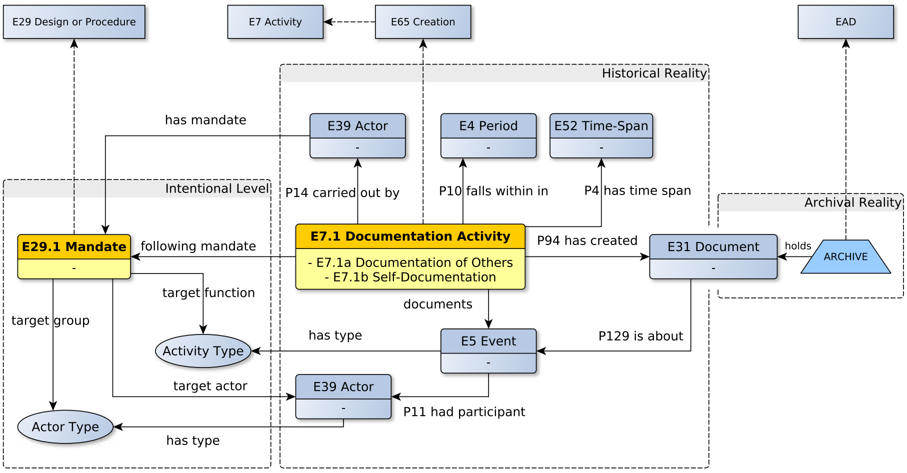


Fig. 5. The current draft of the Documentation-Activity pattern

of actors, or types of activities may be the target. In order to appropriately describe these target relations new properties – *target function*, *target group*, *target actor*, and *has mandate* – have been introduced.

The *E7.1 Documentation Activity* and the *E29.1 Mandate* related through *follows mandate* constitute the essential core of the identified common pattern: The documentation of events according to standing mandates producing documents which can be found in the archive. This mandate-based documentation (*auftragsgemäße Dokumentation*) can not be adequately represented with *E65 Creation* and *E29 Design or Procedure*. The pattern allows to draw conclusions on the probability that specific types of events have been documented and that traces can be expected in the archive.

The documentation activity and its contextual classes can be seen as being part of a description of the *historical reality* as given in the user's question. The mandate, on the other hand, belongs to an *intentional level* (*Absichtsebene*) where principles are formulated which are meant to formally govern the historical reality and which might find their expression in documents. These documents are the point where this ontological representation of the historical reality would intersect with the one of the *archival domain* of record keeping as indicated in figure 5. It is important to note that an *E31 Document* is not a physical item but "comprises identifiable immaterial items that make propositions about reality". A physical materialization of an *E31 Document* in the archive may be an *E33 Linguistic Object* which "comprises identifiable expressions in natural language or languages". Here, a model of expressions of documents in the archive is not included.[13]

[13] Cf. [18] for an approach to mapping EAD to the CIDOC CRM.

The analysis of the questions is on-going and changes to the pattern might occur and there are other aspects which appear to be relevant. The *official* and *unofficial* nature of a document, for example, seems to be another important aspect. This point cannot be discussed in any detail in this paper, however, if a document is official or unofficial is most likely determined by the circumstances of its publication. As already mentioned, the examples also show cases in which the documentation activity is carried out by the same actor who is also responsible for the documented activity. This is a kind of self-documentation giving an "official account" (*Rechenschaftsbericht*) such as proceedings, government statements etc. In the Documentation Activity pattern this can be expressed by two principle sub-types *E7.1a Self-Documentation* and *E7.1b Documentation of Others.*

7 Conclusion

This paper introduced the draft of the *Documentation-Activity* pattern which is part of the *Archival Knowledge Model* (AKM). The AKM is an ontological model which comprises representations of general patterns found in archival user inquiries and their interpretations.

Such an ontological model can help to bridge the semantic gap between traditional archival documentation and organizing principles and the conceptualizations employed by different kinds of users and support building search and discovery systems which are able to better respond to pattern-oriented questions. As a formal model, the AKM could also inform the design of archival metadata schemas or new archival "cataloging rules" as, for example, that titles of series or files should not be plain text but structured according to patterns like the Documentation-Activity pattern.

References

1. Unsworth, J.: Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? (2000)
2. Menne-Haritz, A.: Access: The reformulation of an archival paradigm. Archival Science **1** (2001)
3. Cox, R.: Revisiting the archival finding aid. Journal of Archival Organization **5**(4) (2008)
4. Craig, B.: Perimeters withfences?or thresholds with doors? two views of a border. American Archivist **66**(1) (2003)
5. Sinn, D.: Room for archives? use of archival materials in no gun ri research. Archival Science **10**(2) (2010)
6. Anderson, I.G.: Are you being served? historians and the search for primary sources. Archivaria (58) (2004)
7. Doerr, M., Iorizzo, D.: The dream of a global knowledge network: A new approach. Journal on Computing and Cultural Heritage **1**(1) (2008)
8. Dworman, G.O., Kimbrough, S.O., Patch, C.: On pattern-directed search of archives and collections. Journal of the American Society for Information Science **51**(1) (2000)

9. Shaw, R.: Information organization and the philosophy of history. Journal of the American Society for Information Science and Technology **64**(6) (2013)
10. Harris, C.: Archives users in the digital era: A review of current research trends. Dalhousie Journal of Interdisciplinary Management **1** (2005)
11. Duff, W.M., Johnson, C.A.: A virtual expression of need: An analysis of e-mail reference questions. American Archivist **64**(1) (2001) 43–60
12. Collins, K.: Providing subject access to images: A study of user queries. American Archivist **61**(1) (1998)
13. Conway, P.: Partners in Research: Improving Access to the Nation's Archive. Archives & Museum Informatics, Pittsburgh (1994)
14. Gagnon-Arguin, L.: Les questions de recherche comme matriau dtudes des usagers en vue du traitement des archives. Archivaria **46**(1) (1998)
15. Case, D.O.: The collection and use of information by some american historians: A study of motives and methods. The Library Quarterly **61**(1) (1991)
16. Gardin, J.C.: Archaeological discourse, conceptual modelling and digitalisation: An interim report of the logicist program. The Digital Heritage of Archaeology: Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 30th Conference, Heraklion, Crete, April 2002, CAA 2002 (2002)
17. Doerr, M.: The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. AI Magazine **24**(3) (2003)
18. Bountouri, L., Gergatsoulis, M.: Mapping encoded archival description to CIDOC CRM. Proceedings of the First Workshop on Digital Information Management (2011)

Quality management of 3D cultural heritage replicas with CIDOC-CRM

Nicola Amico[1], Paola Ronzino[1], Achille Felicetti[1], Franco Niccolucci[2]

[1]PIN, VAST-LAB, Prato, Italy
{nicola.amico, paola.ronzino, achille.felicetti}@pin.unifi.it
[2]PIN, VAST-LAB, Prato, Italy
franco.niccolucci@unifi.it

Abstract. The paper proposes to use CIDOC-CRM and its extension CRMdig to document the planning and execution of 3D models of cultural objects in order to manage the quality of the replicas. Full documentation of every process is key to guarantee the quality of the outcomes according to the industrial approach to quality known as Quality Management, for example as described to ISO9001:2008.

Keywords. CIDOC-CRM extension, Quality Management, 3D replicas, cultural heritage

1 Introduction

The use of visual aids to model cultural heritage, besides textual description, has always accompanied the design, planning, creation and documentation of monuments and artifacts. Recently, 3D models are increasingly used thanks to the visualization capabilities of computers and the availability of high-performance graphic cards. A further push to the adoption of 3D models comes from the diffusion of technologies like 3D scanning and photogrammetry that make 3D modeling a widely available methodology. Nowadays it is being adopted for mass acquisition of artifacts and monuments and 3D datasets are stored in an increasing number in openly accessible digital libraries. For example, there are projects aiming at populating Europeana, the European digital library, with 3D models of European art, archaeology and architecture masterpieces or creating tools for the creation of collections of digital replicas of cultural objects [1, 2]. However, issues have been raised about the quality of the 3D models and their suitability to become a substitute of the original, leading to the statement of widely accepted general principles [3]. An engineering approach to quality is based on the quest for details and accuracy and measures quality in microns (model resolution) and number of polygons (level of detail: LOD).
This approach is technology-driven and does not take into account the customers' requirements and perspective. It is also cumbersome to implement, because it requires ex-post verification of the model. Finally, it does not take into account the data acquisition conditions that might adversely influence the model quality, regardless of its

pretended precision. Some institutions in recent years started to define guidelines for a correct use of 3D laser scanner for cultural heritage [4, 5, 6, 7]. The idea behind this approach is that if the acquisition is done “at best”, the result can only be good. This is correct, but how can a subsequent user know it and trust the model? As regards 2D, for instance, it is suggested [8] that direct inspection is carried out either on all the models or on a sample of them – what is clearly unfeasible in the case of complex 3D models.

In a way similar to industry standards, for example ISO9001:2008, a better approach should consider the entire pipeline of 3D model production and document the entire workflow. This will not produce ‘good’ models by itself, but it will produce consistent models and enable users to assess their trustworthiness and suitability for purpose, thus enabling re-use. Documentation is crucial to this model, and a suitable documentation system is – as far as we know – still unavailable. CRMdig [9] marked a significant step towards this goal by extending the well-known CIDOC-CRM to digital matters. In the present paper we propose a draft documentation system for the production of 3D models using laser scanners, based on CIDOC-CRM and its extension CRMdig. Other technologies to create 3D models will follow shortly. A similar approach has been proposed and adopted, in a simplified way, by the already mentioned 3D ICONS project [1]. Experience gained on 3D scanning highlighted issues on the procedures adopted, which can vary in relation with the chosen artifact. Indeed each object has to be scanned following special pipeline related to the object features. Our system considers all the steps of the design and creation of the model until it can be released for further processing or direct use “as is” and this procedure has been tested in a number of archaeological artifacts with a satisfactory result. A similar approach has been pro-posed and adopted, in a simplified way, by the already mentioned 3D ICONS project [1]

2 The scanning workflow

The laser scanner workflow consist of a number of steps, some of which need to follow a precise order. They are:

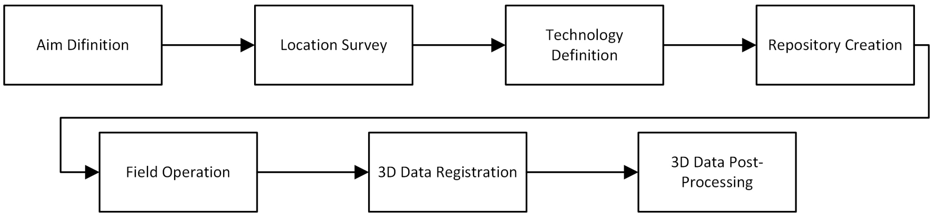
- Aim definition: this step is preliminary and aims at defining the purpose of the digitization. For example, this could be ‘modeling for cultural documentation’, ‘production of models for dissemination’, ‘creation of 3D models for virtual restoration’ and so on.
- Location survey: here a reconnaissance is carried out. The location where the scanning will be performed is surveyed, analyzing the environmental conditions (lighting, temperature, presence of dust, etc.), the features and size of the object compared with the device to be used for the work, and the ‘scene’, i.e. the background surrounding the object to be scanned, for instance, in the case of a monument, the location where it is placed; for a museum object, the space available for scanning, etc. The information recorded includes notes, pictures, sketches, measurements (e.g. of light) and so on. Among others, this stage will support defining the best time to collect the data, identify the presence of highly reflective surfaces, obstruc-

tions and obstacles that may cause voids and artifacts. In outdoor areas, it will be necessary to check weather for rain, fog, dust, heat radiation, which may influence not only the equipment set-up and functioning, but also the outcomes, increasing artifacts and noisy data, and the scanning effective range.

- Technology definition: this step concerns the decision about the device and the technology to be used. Sometimes this choice is dictated by external considerations, as budget or availability. However, the features of the planned scanning may suggest choosing a device and/or a technology instead of another one, so this step interacts with the previous one. For example, the operator may choose among Time-of-Flight (TOF) scanners for long-range acquisition, Phase-based scanners for short-range acquisition, and Triangulation ones for small and medium-sized objects.
- Repository design and creation: in this step the repository is designed according to the project needs. The project may use an existing repository, if the work concerns models that are added to previously existing ones.
- Field operations: this step includes defining the scan position and resolution, the type and number of marks/targets and their position. Each scanner position and orientation angles must be defined according to a local or global site coordinate system. Indoor areas or places (caves, museums etc.) may require the set-up of a lighting system, so the position of every light must be decided and recorded, especially when RGB capture is expected. Some scanners are provided with a built-in digital camera; others use an external digital camera that must be set too. Depending on the object, marks are placed on the object to support the subsequent step called registration. An optimal choice of the marks as regards type (paper, spherical, cylindrical, retro-illuminated and prism) and an accurate recording of their position (using a GPS and/or a Total Station) are crucial to accuracy, as is the scanner Field-of-View (FOV) which together with the object size determines how many scans are taken and need to be registered. Carrying out field operations will follow the design described above. Any change from the planned modality needs to be recorded.
- 3D data registration: as usually it is not possible to scan the object in one scanning step, the separate models obtained with scanning must be assembled in one complete model, availing of common parts which are made to coincide. These may consist in marks placed on the original as easily recognizable points, or images of the object [7]. The registration process also uses the scanner position, previously recorded, or reconstructed using three Ground Control Points (GCP), with the so-called 'indirect registration' [10]. Registration may also be performed without marks (so-called cloud-to-cloud-based registration), but usually this procedure reduces the accuracy of the overall dataset. A pre-registration cleaning is carried out, cleaning the range maps from noisy data and cleaning the borders of each scan, affected by the error of incidence of the laser beam on the surface (mixed edge effect). The parameters of this cleaning stage must be recorded as well.
- 3D data post-processing: this includes all the final operations carried out on the model. The outcomes of the registration process are used as point cloud to generate different outcomes, or processed with different software. After registration, the

point cloud is used to generate a polygonal mesh, by connecting the points in order to create a surface. Before, the point cloud needs to be edited for meshing. Cleaning filters are applied to the point dataset in order to clean up all the noisy and redundant information and edit RGB color. Overlap reduction is also used to move the range maps for a better registration. All these process can be done both manually and automatically. For the creation of the polygonal mesh the Poisson Surface Reconstruction [11] and the Delaunay Triangulation [10] are two of the most common algorithms used to create triangulated meshes from point clouds. All the processing is based on parameters chosen by the operator. Finally, decimation and resampling, particularly suited for 3D model visualization on the web, may be applied, creating a lower resolution model. RGB editing and texture mapping is the final step of the pipeline in order to obtain a photorealistic 3D model.

The above-described pipeline is represented in the diagram below.



3 Documenting the planned production workflow.

In this section we will outline the documentation system of the abovementioned pipeline using CRMdig. The current version is still a draft, testing it in a number of practical examples. Codes in parentheses refer to entities (E) and properties (P) of CIDOC-CRM; while (D) and (L) refer to CRMdig. The overall digitization project is modeled as a D28 Digital Documentation Process consisting of different activities, those forming the production workflow. The diagrams below describe each activity separately, those represented with a dotted border being referred elsewhere in the model.

3.1 Aim Definition

The step is modeled as the creation (E65) of a document (E31) documenting the digitization aim definition, with the participation (P11) of users (E39).

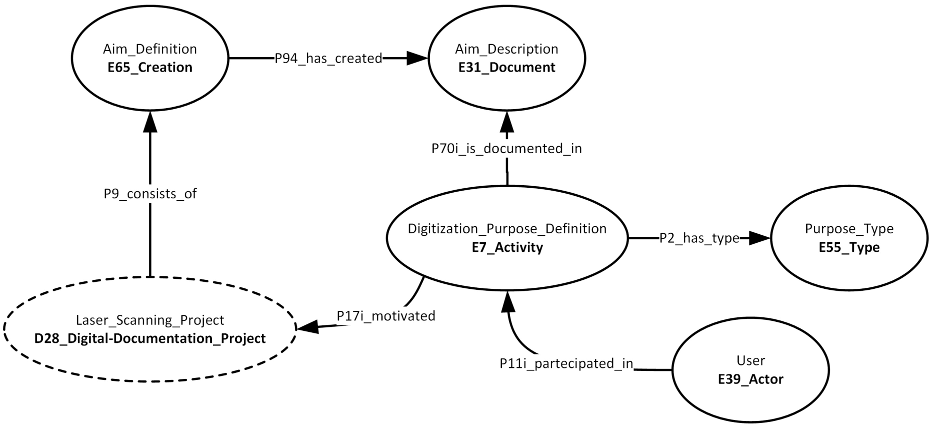


Fig. 1. Aim Definition

3.2 Location Survey

The Location Survey is modeled as an activity (E7), influencing (P15) the choice of the technology to be used. This activity consists of (P9) the Inspection of the Object, of the Site and of an Assessment of the Site Conditions. The Object Inspection is a Creation (E65) of a Document (E31) documenting the Object (E19 Physical Object) to be digitized. The Site Inspection also is a Creation (E65) of a Document (E31). The Place (E53) where the survey takes place (P7) is the same where the Object and its surroundings – the 'scene' – is located. The location property P54 has been chosen because it is intended that the scene is a sort of immovable background. The last component of the survey activity is the Assessment (E13 Attribute Assignment) assigning (P141) a Condition State (E3) to the scene via a P44 condition property.

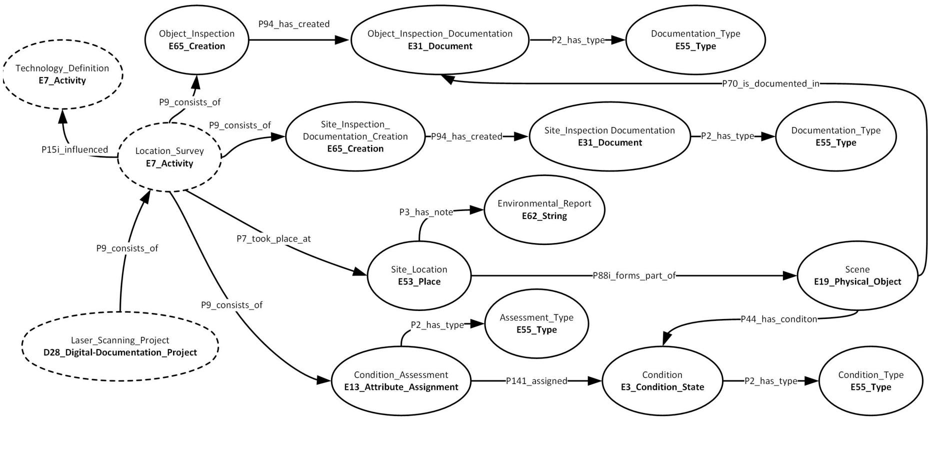


Fig. 2. Location Survey

3.3 Repository Creation

The step consists in the design and creation (P9) of the Repository (D13 Digital Information Carrier) storing the models (D15.Repository Object).

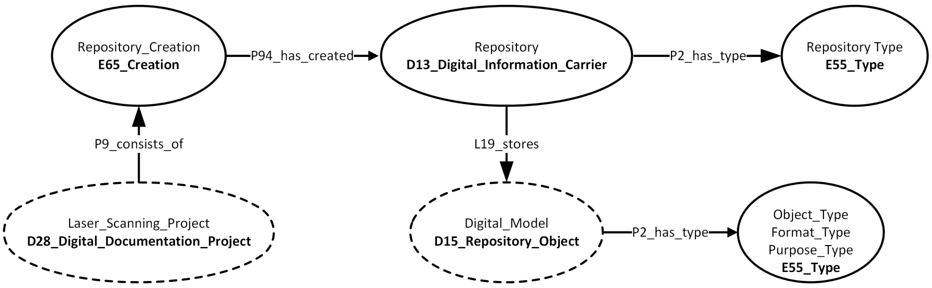


Fig. 3. Repository Creation

3.4 Technology Definition

This step consists of several sub-steps, addressing the different devices to be used in the digitization. It also includes, as specific purpose (P20), the Data Capture Designing (E65), creating (P94) the Digitization Plan (E29).

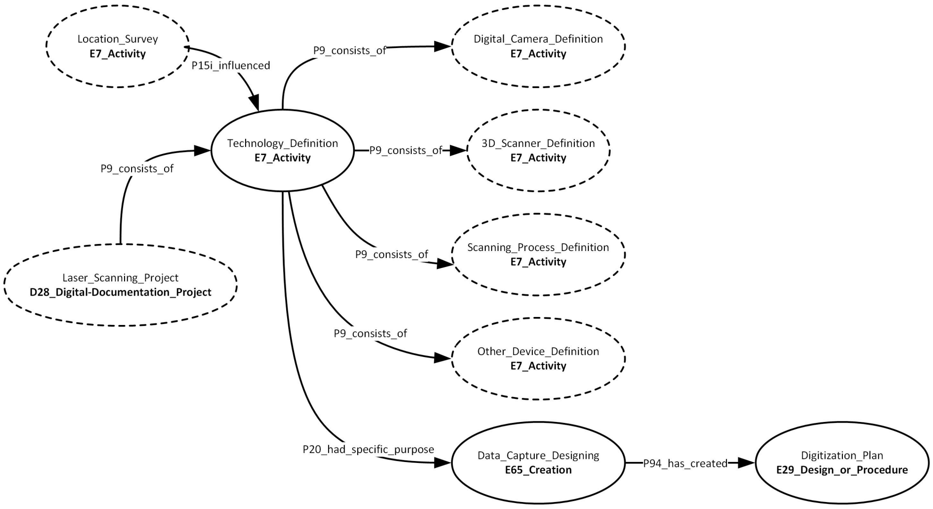


Fig. 4. Technology Definition

Digital Camera Definition. The camera settings are the parameters used (L13) in the Capture Event (D7 Digital Machine Event), altogether considered as a Digital Object (D1), with the values documented via the Event's Dimension (E54). The camera (D8

Digital Device) type collects all its features, and the lenses (E22) type, incorporating their features including the focal length.

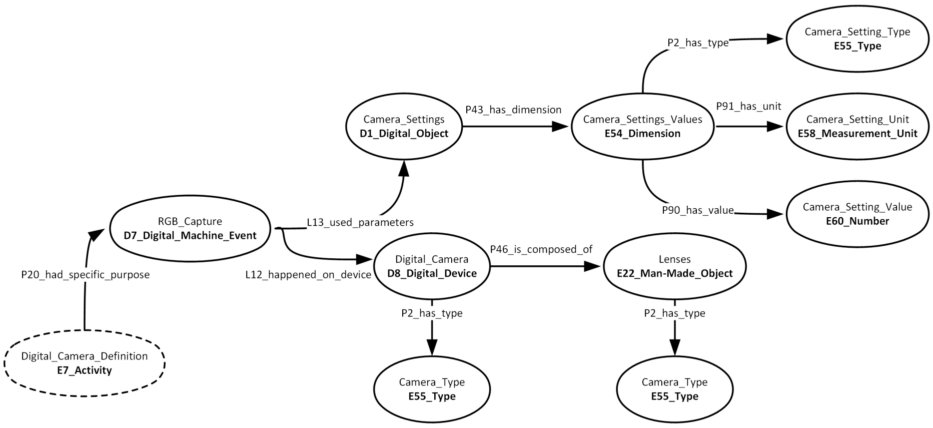


Fig. 5. RGB Capture

Scanner settings. The 3D data capture is modeled in a very similar way, through a Data Capture (D7) Event, which used (L13) parameters (D1) having a Dimension (E34) storing all the necessary information. The type of the Scanner (D8 Digital Device) on which the digitization happens (L12) is recoded separately.

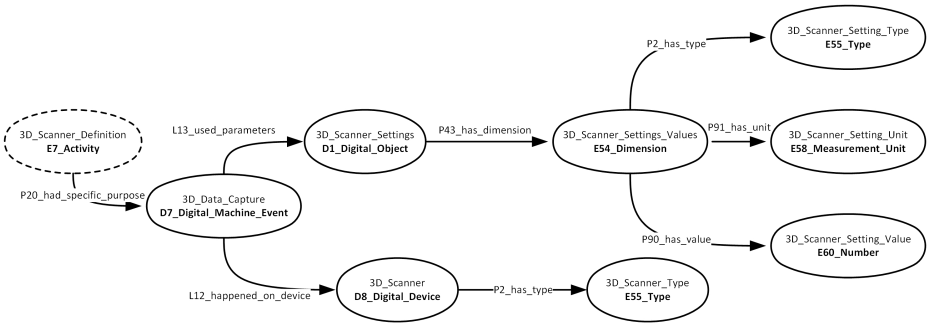


Fig. 6. Scanner Settings

Other Devices. Other devices include equipment for georeferencing the scene and the markers, as a GPS and a Total Station. The structure of their information is very similar to the scanner one and is omitted for space reasons.

3.5 Field Operations.

Field operations concern the creation of a reference network of Marks (E19). Their placement is motivated (P17) by the Scanning Procedure (E29) and the position is recorded through the Measurement (E16) of their Spatial Coordinates (E47).

3.6 Registration (Complete Registration)

Recording the parameters used in the Complete Registration process, modeled as a D7 Digital Machine Event, concerns both the Pre-Registration Cleaning (E9 Formal Derivation) that picks (L21) a model from the (D15) Repository and returns (L22) it there after processing; and the Registration (D10 Software Execution) that takes in input (L10) several models from the Repository (D15) and outputs there (L11) the assembled model. Parameters are modeled as Digital Objects (D1), stored via an E54 Dimension and the related type/unit/value as in the previous cases.

3.7 Post-processing

Post-processing is modeled as a D3 Formal Derivation that picks (L21) a Model from the Repository (D15) and returns (L22) it back after processing. It uses (L23) some software (D1) with has (L13) settings and parameters modeled as usual via E54 Dimension and then type/unit/value.

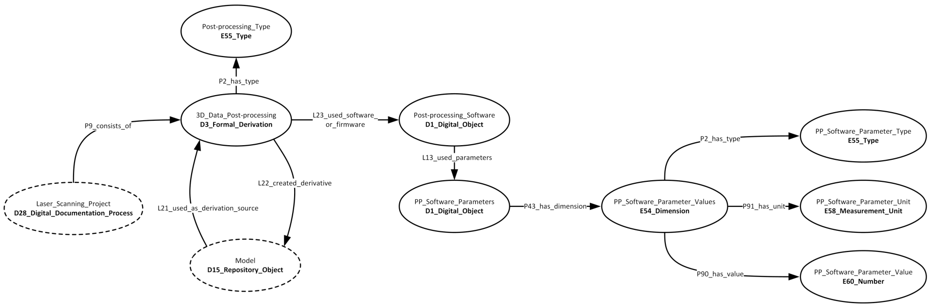


Fig. 7. Post-processing

4 Conclusions and Further Work

With the present paper we have explored how the CRM may support Quality Management, and the conclusion is encouraging. The proposed model may need revision and refinement dictated by practice and perhaps may suggest the definition of shortcuts, such as a simpler way to assign values to parameters. Implementation will need tools to simplify the work.

The CRM had, in the past, the bad reputation of being complicate mainly because of the lack of comfortable input tools for systems based on it. Initial experiences with scanner operators have shown a sort of annoyance for recording all these data. As already noted in 3D COFORM, equipment producers are instead to blame, because they do not provide any information about the device settings, as is done, for instance, in the EXIF file for 2D data capture. Nevertheless, many of the recording tasks may be easily automated designing an intelligent input interface.

5 Acknowledgements

This paper has been partially supported by the European Commission through the ARIADNE project.

References

1. 3D ICONS: http://www.3dicons-project.eu
2. 3D-COFORM: http://www.3d-coform.eu
3. The London Charter: http://www.londoncharter.org
4. Bryan P., Blake B., Bedford J., 2009, Metric Survey Specifications for Cultural Heritage, English Heritage Publishing.
5. Barber D., Mills J., Andrews D., 2011 3D Laser Scanning for Heritage. Advice and guidance to users on laser scanning in archaeology and architecture, English Heritage Publishing, Swindon.
6. Barber D., Mills J., Bryan P., 2004, Towards a standard specification for terrestrial laser scanning of Cultural Heritage, in XXth ISPRS Congress: Proceedings of Commission V, Istanbul, Turkey.
7. Hiremagalur J., Yen K. S., Akin K., Bui T., Lasky T. A., Ravani B., 2011, Creating Standards and Specification for the Use of Laser Scanning in Caltrans Projects, AHMCT Research Center, Davis, USA.
8. Europeana Regia D5.3 Quality Management, pp. 29-31. http://www.europeanaregia.eu/en/project-europeanaregia/project-documentation
9. Doerr, M., Theodoridou, M.: CRMdig: A Generic Digital Provenance Model for Scientific Observation. In: TaPP 2011, Heraklion (2011)
10. Lerma J. L., Van Genechten B., Heine E., Quintero M. S., 2008, 3D RiskMapping – Theory and practice on Terrestrial Laser Scanning, Editor Universidad Politecnica De Valencia, Valencia, Spain.
11. Kazhdan M., Bolitho M, Hugues H., 2006, Poisson Surface Reconstruction, Symposium on Geometry Processing, Cagliari, Italy, pp. 61-70.

European standards for the documentation of historic buildings and their relationship with CIDOC-CRM

Paola Ronzino[1], Nicola Amico[1], Achille Felicetti[1], Franco Niccolucci[2]

[1]PIN, VAST-LAB, Prato, Italy
{paola.ronzino, nicola.amico, achille.felicetti}@pin.unifi.it
[2]PIN, VAST-LAB, Prato, Italy
franco.niccolucci@unifi.it

Abstract. Integration of architectural datasets concerning historic buildings depends on their interoperability, which has as first step a mapping to a common schema. The paper investigates current approaches and proposes mapping to a CIDOC-CRM extension as the common glue to overcome the fragmentation of datasets provided by large national institutions such as MIBAC in Italy, EH in the UK, and so on, and by EU projects, each one structured according to a different metadata schema. The paper describes the mapping of the MA-CA MIBAC-ICCD schemas, probably the most comprehensive, to CRM.

Keywords. CIDOC-CRM, historic buildings

1 Introduction

There is a clear need in Europe of harmonizing actions on built heritage to face the challenges posed by environmental hazards and societal changes. The most comprehensive initiative on this regard is the EU Joint Programming Initiative on Cultural Heritage and Global Change [1], a framework within which EU Member States jointly address areas where public research programmes can respond to major societal challenges concerning heritage and its preservation. The theme has been addressed also by the EU project EU-CHIC (Cultural Heritage Identity Card) [2], which defined the concept of the CHICEBERG Protocol for the integrated documentation of built heritage, based on a taxonomy of historic buildings developed by the EU project Perpetuate [3]. EU-CHIC mainly concerns the conservation and documentation of environmental changes affecting built heritage assets, such as historic buildings and monuments. Most countries in Europe have developed their own systems for storing information concerning built heritage: among others, the Italian Ministry of Culture MIBAC that adopts forms prepared by a specialized institute (ICCD, Central Institute for Cataloguing and Documentation [4]); English Heritage, using the MIDAS scheme [5]; the French Ministère de la Culture, using the Schéma Documentaire Appliqué au Patrimoine et à l'Ar-

chitecture (SDAPA) [6]. Moreover, European projects contributing to Europeana, the European digital Library, have developed their own schemas and mapped them to EDM, the Europeana Data Model. Such projects include CARARE [7] and 3D ICONS [8]. In conclusion, there is a number of different metadata schemas organizing large datasets but preventing any effort for dataset integration, which is an absolute need to develop European policies for research, conservation, restoration and dissemination. Such datasets intersect those considered by ARIADNE [9], the European Research Infrastructure for archaeological datasets, as far as built heritage includes archaeological remains. ARIADNE aims at providing an integrated access to archaeological datasets throughout Europe, and is developing an extension of CIDOC-CRM to guarantee their interoperability [10]. It seems therefore that CIDOC-CRM, or if necessary an extension of it, is the key to overcome the fragmentation of architectural datasets, and this is the way we propose to follow. We are currently building a mapping from each of the metadata schemas used in the most important European repositories, such as those mentioned above, i.e. the ICCD schemas, MIDAS, CHICEBERG and the CARARE/3D ICONS schemas, to the CIDOC CRM. It is a complicated work, because it involves more than 700 fields, some identical in meaning, some just similar but with a different nuance, and other very different. A preliminary version of the mapping is ready and will be published on VAST-LAB's web site [11]. The mapping of the CARARE schema to CIDOC CRM has been discussed in [12].
In our experience, the most comprehensive is the ICCD one, and we are working closely with the Institute to develop the mapping of the many forms it uses. A full description of the forms may be found on the ICCD site [4].In this paper we will present a draft mapping of the ICCD Monument form to CIDOC CRM; or, better, an outline of it, for space reasons. The full version is going to be available on the above-mentioned VAST-LAB's web site as well.

2 The ICCD MA/CA form

The MA/CA form is used for archaeological monuments and complexes [13]. As regards architecture, there is a similar form called form A [14], used for historic buildings, which has only slight differences from MA/CA. We have mapped both, but for the sake of brevity we will present here only the MA/CA to CRM mapping. The MA/CA form includes more than 300 fields, each identified by a unique letter code and a name. We will use only the code and give an informal English translation of the name. Metadata are grouped in the following 'wrappers': CD-AC – Codes; RV – Relationships; OG – Object; LC – Current Location; CS – Cadaster; LS – Historic Location; GP-GL-GA – Georeferencing; RE – Way of discovery; DT – Chronology; AU – Cultural definition; RO – Reuse; MT – Technical data; CO – Conservation; RS – Res-

toration; DA – Analytical Data; MC – Samples and analyses; TU – Legal status; DO – Sources; AD – Data access; CM - Compiler; AN – Notes.
Fields (and wrappers) of little interest for integration will not be considered.

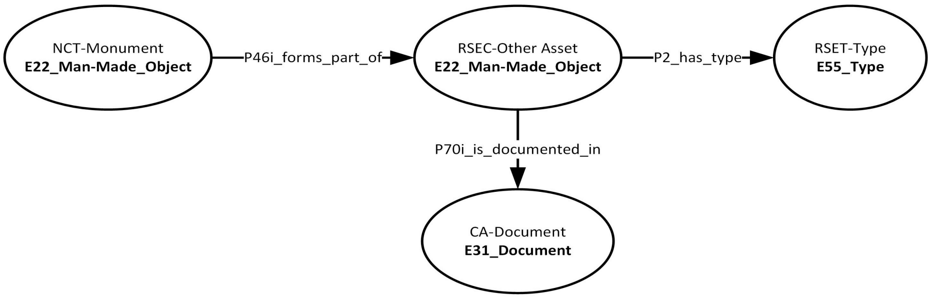
3 The mapping

3.1 RV – Relationships

This set of fields is used to document the relationship of the monument, identified with its unique code NCT, with other assets of different kind. In the relationships below, the domain is the monument and the range is the other asset, which can belong to the same category or can be different. Entities corresponding to MA/CA fields are identified with the MA/CA letter code.

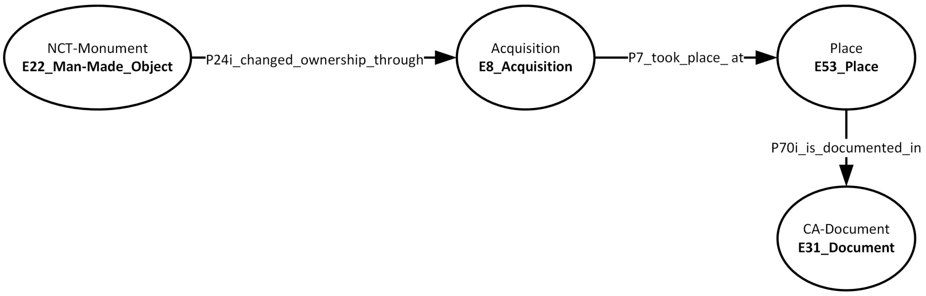
- Is contained in:

The monument relates to another monument (MA) or archeological complex (CA), which represents the monument location at the time of cataloguing.



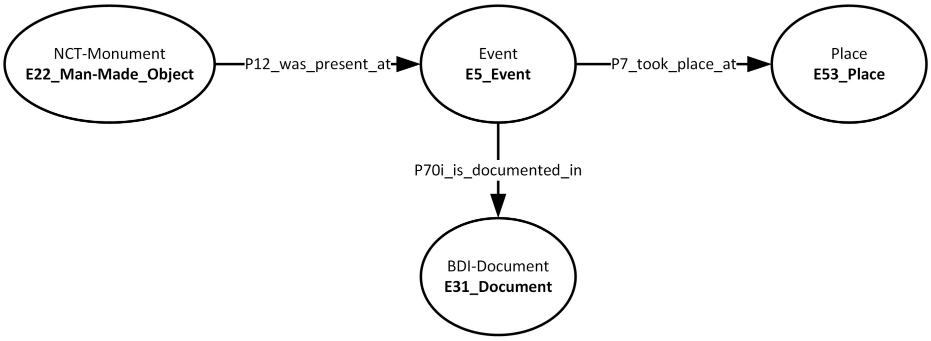
- Was found in:

This relation links the monument (MA) or archaeological complex (CA) to the site (SI form) or Stratigraphic Essay (SAS form) where it was found.



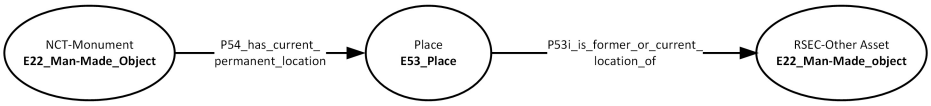
This path is not completely convincing and perhaps a better way of documenting archaeological discovery could be considered in a future extension of CIDOC CRM.

- Is involved in:

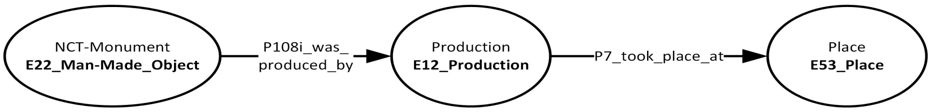


This documents the connection between the monument, and an event (such as a festivity, celebration, rite, etc.), documented in a form pertaining to intangible heritage.

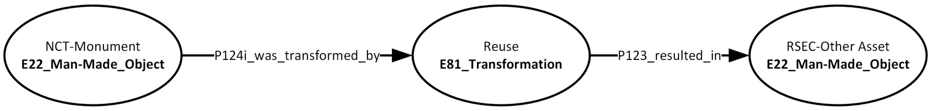
- Has environmental/spatial relationships with:



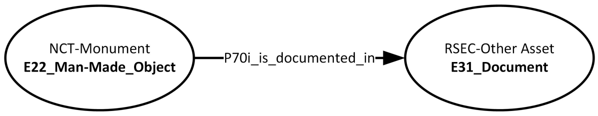
- Was made in:



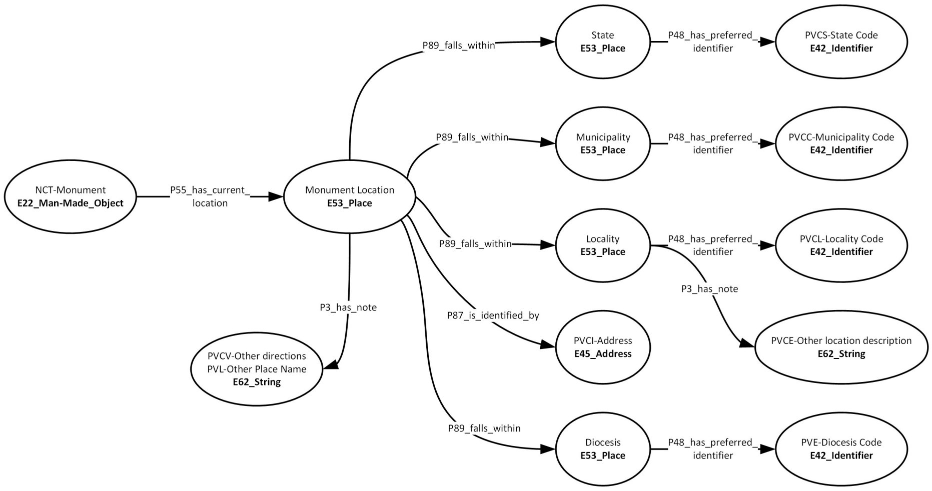
- Is reused by:



- Is documented in:



3.2 LC – Current Location

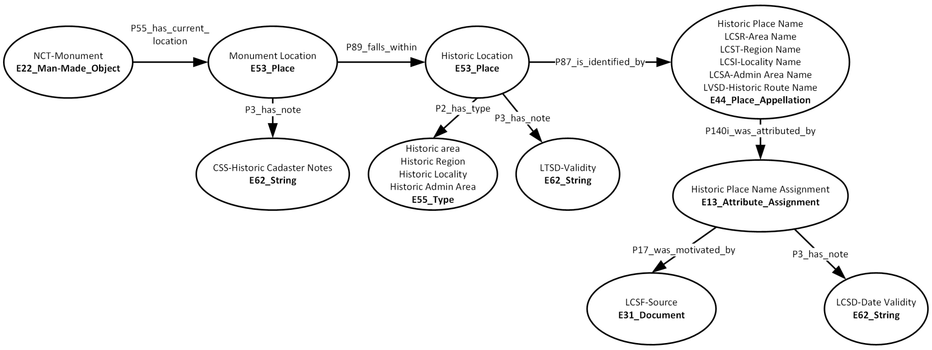


As shown by the diagram above, metadata about location are modeled via the monument location (E53) that falls within (P89) various other places useful to define the location.

3.3 LS – Historic Location

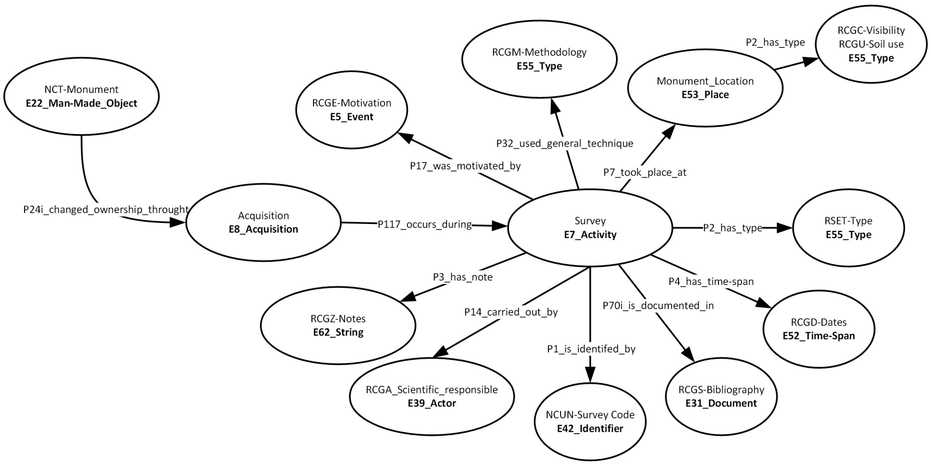
Historic Location relates the monument to various historic places, such as areas, roads and places with their place names. This correspondence is modeled via the Monument Location, as before, which receives (P140i) by an Attribute Assignment (E13) the assignment of various historic locations (E53) with their place names or other specification (E44 Place Appellation).

We used an E62 String to express the time validity of the historic reference as a note to the Historic Place Name assignment, since CIDOC-CRM does not seem to have a simple way of expressing the time validity of a historic localization.

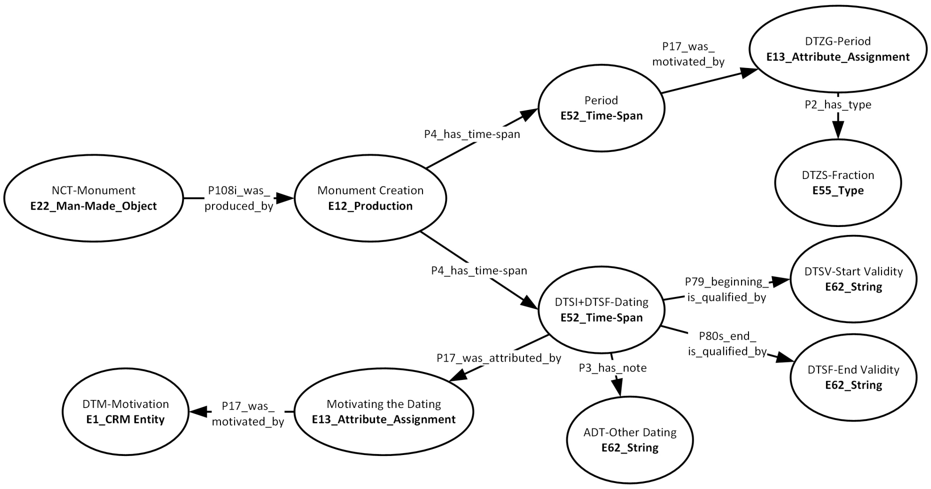


3.4 RE – Way of Discovery

This wrapper collects information about the way the monument was discovered, distinguishing among survey, excavation and other investigations. The diagram below concerns the survey, while the excavation one is very similar. The modeling starts with an ‘Archaeological Discovery’, on which the same comments as above can be made. In this case it occurred during a Survey (E7) Activity, identified by its code NCUN for which – as for any field whose code begins with N – there is an authority file. The Survey took place (P7) at the Monument Location (E53) about which RGCU Soil Use and RCGC Visibility of the terrain are recorded as types (E55). Information about the survey concerns among others its RCGD Date (E52 Time Span), RCGA who did it (E39 Actor), and the Methodology type (E55) used. The reason RCGE for carrying out the survey is modeled as an E5 Event.



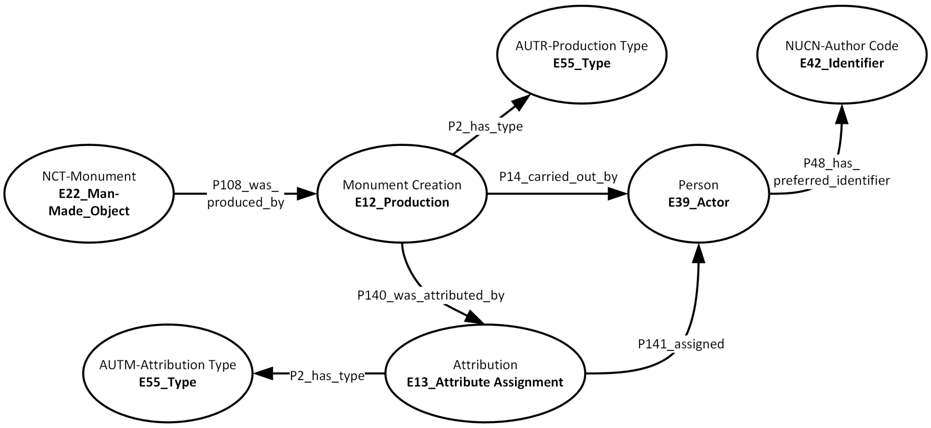
3.5 DT – Chronology



The chronology section is based on an E12 Production event. Chronology may be approximate, falling within the DTZG Period (E52 Time-Span), affected by a qualifier DTZS Fraction, modeled as E55 Type, e.g. ‘end of’, ‘early’, and so on; or more precise, but possibly still approximate such as “ante 1410 AD”, “approx. 600 BC” etc., with a start and an end date incorporated in DTSI+DTSF Dating, an E52 Time-Span, start qualified (P79) and end qualified (P80) by validity, respectively DTSV and DTSL, as ‘ante’, ‘approx.’ etc.

3.6 AU – Cultural Definition

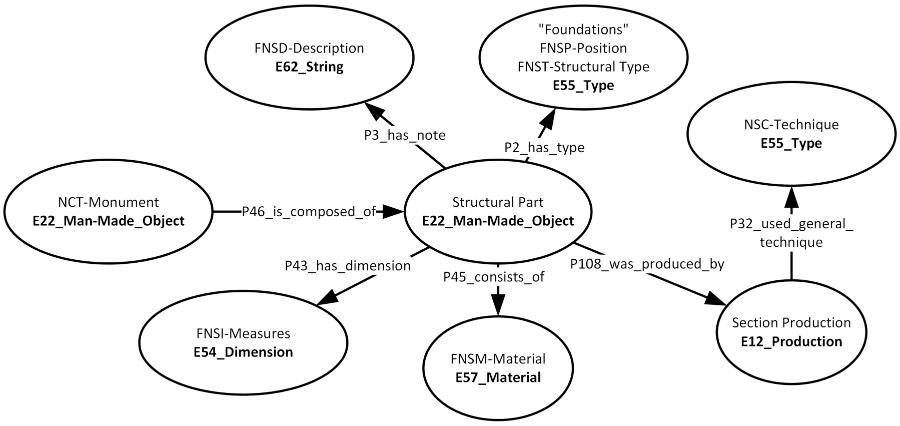
This section concerns authorship, and is centered on the Monument Creation, an E12 Production. The Author is an E39 Actor. It could be an E21 Person identified (P48) by the NUCN Author Code that refers to the AUT authority file, which includes all the information concerning the author. If the identification is imprecise, reference to “school of”, “workshop of”, or “group of” is included in a special field called AUTS. These special cases lead to slightly different modeling (not presented here for space reasons), where the Author is an E74 Group and the participation of a person in this is modeled with P15 was influenced by, for “school of”; P107i is current or former member of, for “group of”; and so on. AUTM, the Motivation of the attribution, is modeled via an E13 Attribute Assignment, which assigns the Author to the Production. The mapping of additional information, sometimes present, concerning the cultural ambit, i.e. generic cultural references to a cultural context, and the commission of the monument is not detailed here for the sake of space.



3.7 DA – Analytical Data

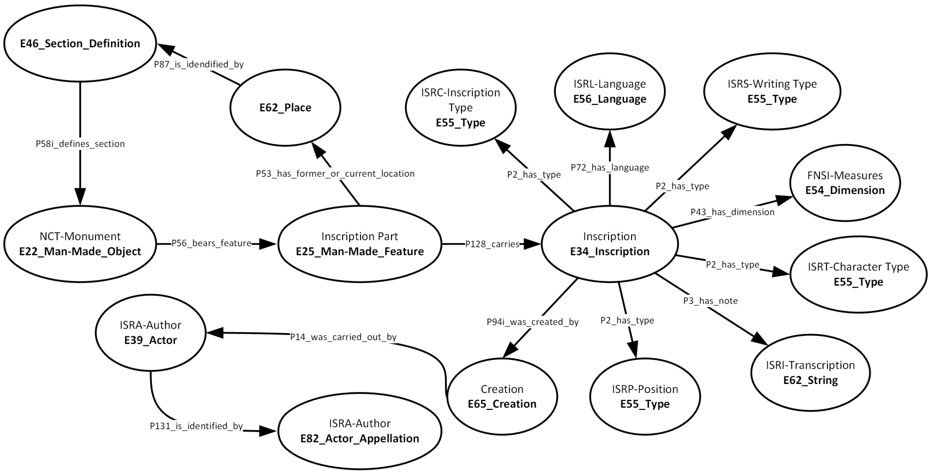
This section describes the structural parts of the monument: foundations, vertical and horizontal structures, stairs, the roof, open spaces, and includes marks, inscriptions and emblems. Each one of these has a separate subsection.

The diagram below concerns foundations. They are modeled as a part of the monument, defined as another E22 Man-Made Object of type “Foundations”. Besides the FNSD Description, modeled as an E62 String, and several types assigned to the part, the information recorded includes FNSM Material, modeled as E57 Material, the material used for the foundations such as bricks, stones, unknown etc.; and the construction technique, modeled via an E12 Production event relating to the part, which used as general technique (P32) the FNSC Technique, an E55 Type. Information concerning horizontal and vertical structures, the stairs, the roof and open spaces is very similar and is modeled in the same way, with more types characterizing the different parts.



3.8 ISR – Inscriptions

The following diagram describes the model for the inscriptions.



The interpretation of the modeling is straightforward. A difficulty here concerns the text and author of the text of the inscription. In some cases the original ISRA Author field contains mixed information, such as the author and the work from which the inscription text is taken, so modeling it as an E62 String is somehow compulsory, as a consequence of overloading the field with too much information in the source data model. But in other cases, if for example only the text author is documented and further elaborated with information on the person, modeling it as a String leads to a cul-de-sac. To provide a more structured and detailed information, whenever possible both author identification and attribution must be described. To identify the author, a path such as E34 Inscription – P94i was created by – E65 Creation – P14 carried out by – E39 Actor – Actor P131 is identified by – E82 Actor Appellation, may be used. If comments on the attribution must be made, e.g. to qualify its reliability or source, this path may be substituted with E34 Inscription – P140 was attributed by – E13 Attribute Assignment (of authorship) – P140 assigned – E39 Actor, and then further qualifying the authorship attribution E13.

4 Conclusions and Further Work

For space reasons, it was impossible to present here a complete description of the mapping, but we hope that the section dealt with gave the flavor of the work. In conclusion, the mapping is feasible and perhaps improves the original documentation scheme without loosing its richness of details. The ongoing mappings of other national repositories of monument documentation, and the creation of multilingual thesauri that are also in progress (see [15] for further details) show that the interoperability of monument datasets is feasible, if not easy, and that the integration of these repositories at European level would create an infrastructure as useful as the forthcoming archaeological one. Further work will concern completing the mapping of other ICCD schemas relating to architecture and addressing conservation and restoration, which are present in these forms in a very succinct way.

5 Acknowledgements

The present work has been partially supported by the European Union through the ARIADNE project.

References

1. JPI: `http://www.jpi-culturalheritage.eu`
2. CHICEBERG `http://eu-chic.eu/index.php/news/entry/chiceberg/`
3. Perpetuate: `http://www.perpetuate.eu`
4. ICCD: `http://www.iccd.beniculturali.it`
5. MIDAS: `http://www.english-heritage.org.uk/publications/midas-heritage/`
6. SDAPA: `http://www.culture.gouv.fr/culture/dp/schemaDAPA/index.html`
7. CARARE metadata schema 2.0, `http://www.carare.eu/eng/Resources/CARARE-Documentation/CARARE-2.0-schema`
8. 3D ICONS: `http://www.3dicons-project.eu`
9. ARIADNE: `http://www.ariadne-infrastructure.eu`
10. Ronzino, P., Niccolucci, F., D'Andrea, A.: Built Heritage metadata schemas and the integration of architectural datasets using CIDOC-CRM. Paper accepted at "Built Heritage" Conference, Milan November 2013
11. Ronzino, P., Amico, N., Felicetti, A., Niccolucci, F.: Built heritage documentation in CIDOC-CRM. PIN technical report http://www.vast-lab.org/documents (2013)
12. Ronzino, P., Amico, N., Niccolucci, F.: Assessment and comparison of metadata schemas for architectural heritage. In Proc. of the XXIII International CIPA Symposium (2011)
13. ICCD Scheda MA-CA Monumento Archeologico Complesso Archeologico v.3.00 (2013) `http://www.iccd.beniculturali.it/index.php?it/251/beni-archeologici`
14. ICCD Scheda A Beni architettonici e ambientali Edifici e manufatti v.3.00 (2013) `http://www.iccd.beniculturali.it/index.php?it/252/beni-architettonici-e-paesaggistici`
15. Ronzino P., Amico N., Niccolucci F. Federating Specialized Digital Libraries. In: Niccolucci F., Dellepiane M., Pena Serna S., Rushmeier H., Van Gool L. (eds) Proceedings VAST2011, Eurographics, pp. 97-103

Large-scale Reasoning with a Complex Cultural Heritage Ontology (CIDOC CRM)

Vladimir Alexiev, Dimitar Manov, Jana Parvanova, Svetoslav Petrov

Ontotext Corp, Sofia, Bulgaria
Email: {first.last}@ontotext.com

Abstract. The CIDOC Conceptual Reference Model (CRM) is an important ontology in the Cultural Heritage (CH) domain. CRM is intended mostly as a data integration mechanism, allowing reasoning and discoverability across diverse CH sources represented in CRM. CRM data comprises complex graphs of nodes and properties. An important question is how to search through such complex graphs, since the number of possible combinations is staggering. One answer is the "Fundamental Relations" (FR) approach that maps whole networks of CRM properties to fewer FRs, serving as a "search index" over the CRM semantic web.

We present performance results for an FR Search implementation based on OWLIM. This search works over a significant CH dataset: almost 1B statements resulting from 2M objects of the British Museum. This is an exciting demonstration of large-scale reasoning with real-world data over a complex ontology (CIDOC CRM). We present volumetrics, hardware specs, compare the numbers to other repositories hosted by Ontotext, performance results, and compare performance of a SPARQL implementation.

Keywords: CIDOC CRM, cultural heritage, semantic search, Fundamental Relations, OWLIM, semantic repository, inference, performance, benchmark

1 Introduction

The CIDOC Conceptual Reference Model (CRM)[1] is an important ontology in the Cultural Heritage (CH) domain. CRM is intended mostly as a data integration mechanism, allowing reasoning and discoverability across diverse CH sources represented in CRM. CRM data comprises complex graphs of nodes and properties. An important question is how to search through such complex graphs, since the number of possible combinations is staggering. The "Fundamental Relations" (FR) approach [2,3] "compresses" the semantic network by mapping whole networks of CRM properties to fewer FRs that serve as a "search index" over the CRM semantic web and allow the user to use a simpler query vocabulary.

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

In [1] we published an implementation of CRM FRs created within the ResearchSpace project[2] using OWLIM[3] [4], and presented preliminary performance results. Here we present a revised implementation, provide volumetrics and hardware specs, performance results over the full CRM repository comprising over 2M CH objects of the British Museum (BM), compare the numbers to other repositories hosted by Ontotext, and compare the performance to one based on SPARQL.

2 Specifics

2.1 Implemented FRs

We implemented the following FRs. Compared to [1] these are adjusted after initial experimentation and gained user experience in RS. Each FR has domain Thing and range indicated in parentheses. rso:E55_Technique is a subclass of crm:E55_Type that we use for focused searching of Techniques. The last 5 FRs (17-23) are special extensions:

1. rso:FR92i_created_by (crm:E39_Actor): Thing (or part/inscription thereof) was created or modified/repaired by Actor (or group it is member of, e.g. Nationality)
2. rso:FR15_influenced_by (crm:E39_Actor): Thing's production was influenced/motivated by Actor (or group it is member of). E.g.: Manner/ School/ Style of; or Issuer, Ruler, Magistrate who authorised, patronised, ordered the production.
3. rso:FR52_current_owner_keeper (crm:E39_Actor): Thing has current owner or keeper Actor
4. rso:FR51_former_or_current_owner_keeper (crm:E39_Actor): Thing has former or current owner or keeper Actor, or ownership/custody was transferred from/to actor in Acquisition/Transfer of Custody event
5. rso:FR67_about_actor (crm:E39_Actor): Thing depicts or refers to Actor, or carries an information object that is about Actor, or bears similarity with a thing that is about Actor
6. rso:FR12_has_met (crm:E39_Actor): Thing (or another thing it is part of) has met actor in the same event (or event that is part of it)
7. rso:FR67_about_period (crm:E4_Period): Thing depicts or refers to Event/Period, or carries an information object that is about Event, or bears similarity with a thing that is about Event
8. rso:FR12_was_present_at (crm:E4_Period): Thing was present at Event (eg exhibition) or is from Period
9. rso:FR92i_created_in (crm:E53_Place): Thing (or part/inscription thereof) was created or modified/repaired at/in place (or a broader containing place)
10. rso:FR55_located_in (crm:E53_Place): Thing has current or permanent location in Place (or a broader containing place)

[2] http://www.researchspace.org
[3] http://www.ontotext.com/owlim

11. rso:FR12_found_at (crm:E53_Place): Thing was found (discovered, excavated) at Place (or a broader containing place)
12. rso:FR7_from_place (crm:E53_Place): Thing has former, current or permanent location at place, or was created/found at place, or moved to/from place, or changed ownership/custody at place (or a broader containing place)
13. rso:FR67_about_place (crm:E53_Place): Thing depicts or refers to a place or feature located in place, or is similar in features or composed of or carries an information object that depicts or refers to a place
14. rso:FR2_has_type (crm:E55_Type): Thing is of Type, or has Shape, or is of Kind, or is about or depicts a type (e.g. IconClass or subject heading)
15. rso:FR45_is_made_of (crm:E57_Material): Thing (or part thereof) consists of material
16. rso:FR32_used_technique (rso:E55_Technique): The production of Thing (or part thereof) used general technique
17. luc:myIndex (rdfs:Literal): The full text of the thing's description (including thesaurus terms and textual descriptions) matches the given keyword. FTS using Lucene built into OWLIM.
18. rso:FR108i_82_produced_within (rdfs:Literal): Thing was created within an interval that intersects the given interval or year.
19. rso:FR1_identified_by (rdfs:Literal): Thing (or part thereof) has Identifier. Exact-match string
20. rso:FR138i_has_representation (xsd:boolean): Thing has at least one image representation. Used to select objects that have images
21. rso:FR138i_representation (crm:E38_Image): Thing has image representation. Used to fetch all images of an object
22. rso:FR_main_representation (crm:E38_Image): Thing has main image representation. Used to display object thumbnail in search results
23. rso:FR_dataset (rdfs:Literal): Thing belongs to indicated dataset. Used for faceting by dataset

2.2 OWLIM Rules

We used OWLIM Rules to implement the FRs: a total of 120 rules:

- 14 rules implement RDFS reasoning, a small subset of OWL (owl:TransitiveProperty, owl:inverseOf) and ptop:transitiveOver from the PROTON ontology[4]. These are copied from standard rulesets, as described in [5]
- 106 rules implement FRs. We use a method of decomposing the network of an FR in pieces [1]: conjunctive (e.g. checking the type of a node), disjunctive (parallel), serial (property path), transitive. We implement each piece as a sub-FR and use it to build up bigger pieces.

To deal with the complexity of implementation, we used several approaches:

[4] http://www.ontotext.com/proton-ontology

- A rule shortcut syntax that renders each rule on one line, instead of a line for each premise and conclusion
- A literate programming style, where rule definitions are interspersed with diagrams, discussion and justification in a wiki
- Checking that only known properties and classes are used in the rules (the dependency graph in the next section helped for this)
- Checking that rule variables are used in a linear way (premise variables make a chain, and the conclusion uses the ends of the chain), or in type checks. E.g.

```
x <rdf:type> <rso:FC70_Thing>; x <crm:P46_is_composed_of> y  => x <rso:FRT_46_106_148> y
x <rso:FRT_46_106_148> y; y <crm:P46_is_composed_of> z => x <rso:FRT_46_106_148> z
p <ptop:transitiveOver> q; x p y; y q z => x p z
```

(a) First rule: x is used in a type check, and x-y=>x-y is a linear chain.
(b) Second rule: x-y;y-z=>x-z is a linear chain.
(c) Third rule: p and q are **not** used in a linear way. These variables are in "property" position, and our check skips such variables

A lot more implementation details can be found at the ResearchSpace wiki[5]. The following OWLIM reasoning features[6] were important for the implementation:

- Custom rule-sets. The standard semantics that OWLIM supports (RDFS, RDFS Horst, OWL RL, QL and DL) are also implemented as rulesets.
- Fully-materializing forward-chaining reasoning. Rule consequences are stored in the repository and query answering is very fast.
- sameAs optimization that allows fast cross-collection search using coreferenced values (e.g. Agent URIs)
- Incremental retraction: when a triple is deleted, OWLIM removes all inferred consequences that are left without support (recursively). In order to facilitate this, OWLIM rules have a simple syntax, so they can be checked in "reverse".
- Incremental insert: when a triple is inserted (even an ontology triple), all rules are checked. If a rule fires, the new conclusion is also checked against the rules, etc.
- Efficient rule execution: rules are compiled to Java and executed quickly. For example, we decided late in the game that we want FR45 "Thing is made of Material" to be transitive over the "broader" hierarchy. We added the 2 triples below, and 1M new triples were inferred within 10 minutes (see the implementation of ptop:transitiveOver in (c) above).

```
rso:FR45_is_made_of ptop:transitiveOver skos:broader, crm:P127_has_broader_term.
```

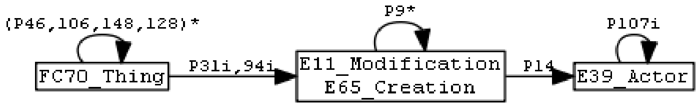
OWLIM rules also have their disadvantages, as described in [1] section 5.3. Chief among them is inflexibility: if the ruleset is changed, the OWLIM server needs to be restarted. Furthermore, if the ruleset should infer different conclusions from the exist-

[5] https://confluence.ontotext.com/display/ResearchSpace/FR+Implementation
[6] http://owlim.ontotext.com/display/OWLIMv53/OWLIM-SE+Reasoner

ing triples, the repository needs to be reloaded. But newly added triples are checked against the rules, as shown in the previous example.

2.3 Example: FR92i_created_by

As an example, let's consider FR92i_created_by "Thing created by Actor", which we define as "Thing (or part/inscription thereof) was created or modified/repaired by Actor (or a group it is a member of)":



This FR includes the following source properties:

- P46_is_composed_of, P106_is_composed_of, P148_has_component: navigates object part hierarchy
- P128_carries: to transition from object to Inscription carried by it
- P31i_was_modified_by (includes P108i_was_produced_by), P94i_was_created_by: Modification/Production of physical thing, Creation of conceptual thing (Inscription)
- P9_consists_of: navigates event part hierarchy (BM models uncorrelated production facts as sub-events)
- P14_carried_out_by, P107i_is_current_or_former_member_of: agent and groups he's member of

This FR uses a previously defined sub-FR FRT_46_106_148_128 (the first loop) and defines another sub-FR:

- FRX92i_created := (FC70_Thing) FRT_46_106_148_128* / (P31i | P94i) / P9*

The sub-FR extends to the Modification/Creation node including the P9 loop and is implemented with 5 rules:

```
x <rdf:type> <rso:FC70_Thing>; x <crm:P31i_was_modified_by> y => x <rso:FRX92i_created> y
x <rdf:type> <rso:FC70_Thing>; x <crm:P94i_was_created_by>  y => x <rso:FRX92i_created> y
x <rso:FRT_46_106_148_128> y; y <crm:P31i_was_modified_by> z => x <rso:FRX92i_created> z
x <rso:FRT_46_106_148_128> y; y <crm:P94i_was_created_by>  z => x <rso:FRX92i_created> z
x <rso:FRX92i_created> y; y <crm:P9_consists_of> z => x <rso:FRX92i_created> z
```

Finally, the FR uses the sub-FR (which also reused in another FR!), and is implemented with 2 rules:

- FR92i_created_by := FRX92i_created / P14 / P107i*

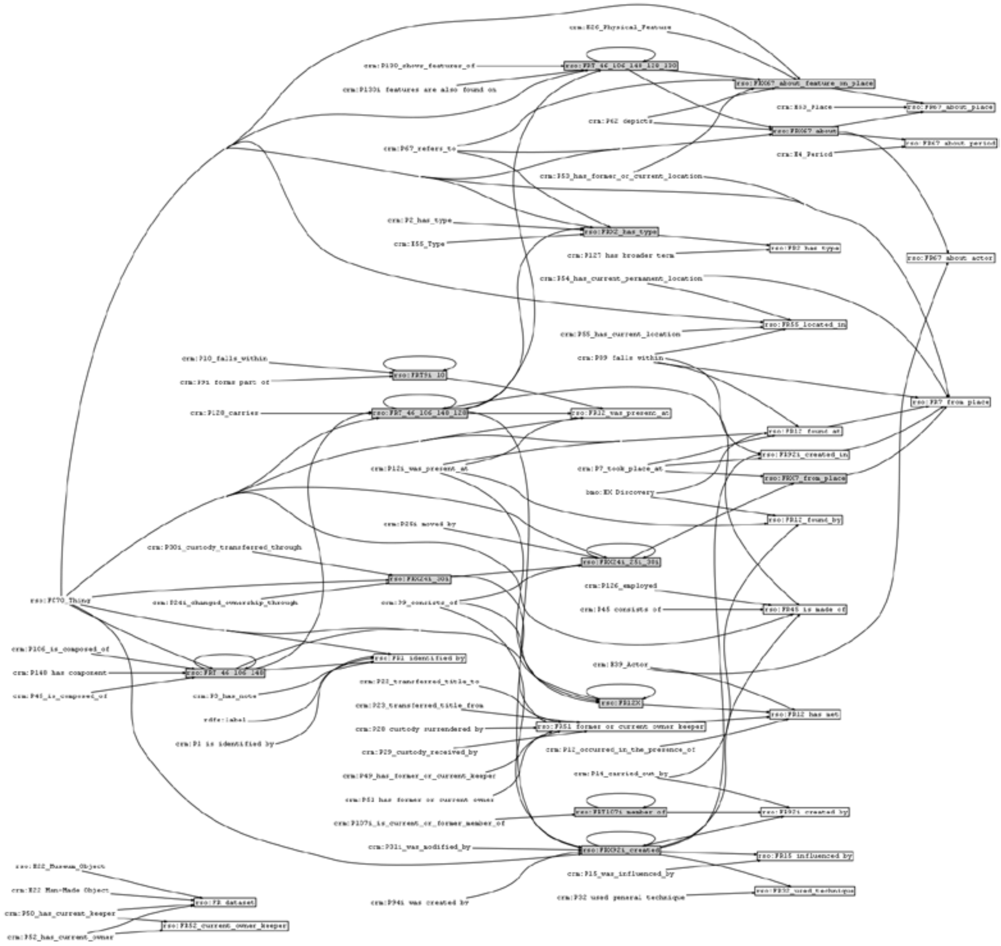
```
x <rso:FRX92i_created> y; y <crm:P14_carried_out_by> z => x <rso:FR92i_created_by> z
x <rso:FRX92i_created> y; y <crm:P14_carried_out_by> z; z <rso:FRT107i_member_of> t =>
  x <rso:FR92i_created_by> t
```

2.4 Sub-FRs and Dependency Graph

The dependency graph of our implementation is shown below, a zoomable version is also available[7]. It has:

- 51 source classes/properties, shown as plain text
- 13 intermediate sub-FRs, shown as filled rectangles. These sub-FRs are used by several FRs to simplify the implementation
- 19 target FRs, shown as rectangles

The diagram illustrates the complexity of the implementation. We used it to verify the implementation as OWLIM rules (e.g. that there are no disconnected properties, each FR uses all source properties as expected, etc).



[7] https://www.dropbox.com/s/6lz48qfdbitkvlk/FR-graph.png

2.5 Hardware Specification

- CPU: Intel Xeon E5-2650 2.00GHz, 20M Cache, 32 cores
- RAM: 128 GB RDIMM, 1600 MHz
- Solid-State Disks: 4*200GB SSD, SATA.
- Hard Disks: 3*300GB, SAS 6Gbps, 2.5-in, 15K RPM.
- Server cost: under $10k.

Large-scale OWLIM deployments are recommended to use SSD for faster disk speed, and the zFS compressing file-system for better SSD utilization and even faster speed. zFS is native for Solaris, but we now have successful deployments on Linux as well. This system has a lot of spare capacity: the hard disks and zFS are currently not used.

2.6 Volumetrics

Some numeric data of our implementation, with discussion:

- Museum objects: 2,051,797 (entities with type rso:FC70_Thing). Most of these are from the British Museum. We are currently completing the ingest of Yale Center for British Art objects into ResearchSpace.
- Thesaurus entries: 415,509 (type skos:Concept). All kinds of "fixed" values that are used for search: object types, materials, techniques, people, places, … (a total of 90 ConceptSchemes)
- Explicit statements: 195,208,156. We estimate that of these, 185M are for objects (90 statements/object) and 9M are for thesaurus entries (22 statements/term).
- Total statements: 916,735,486. The expansion ratio is **4.7x** (i.e. for each statement, 3.7 more are inferred). This is considerably higher compared to the typical expansion for general datasets (e.g. DBpedia, GeoNames, FactForge) that is 1.2 - 2x, and is due to the complexity described below.
- Nodes (unique URLs and literals): 53,803,189. (We don't use blank nodes)
- Repository size: 42 Gb, object full-text index: 2.5 Gb, thesaurus full-text index (used for search auto-complete): 22Mb.
- Loading time (including all inferencing): 22.2h on RAM drive; 32.9h on hard-disks.

2.7 Complexity: Classes

CIDOC CRM is a complex ontology. The deepest branch of the class hierarchy[8] is 10 levels: E1>E77>E70>E71>E28>E90>E73>E36>E37>E34_Inscription. Furthermore, multiple inheritance is used extensively, e.g. E33 is also a super-class of E34_Inscription. For each inscription, 12 type statements are inferred. We use the Erlangen CRM mapping to OWL[9] because it provides inverse and transitive proper-

[8] http://www.cidoc-crm.org/cidoc_graphical_representation_v_5_1/class_hierarchy.html

[9] http://erlangen-crm.org

ties. But it includes a lot of owl:Restriction anonymous classes, e.g. (in Manchester notation)

```
E30_Right SubClassOf: P104i_applies_to some E72_Legal_Object
```

These anonymous classes are useless to us, so we wrote a tool that derives simpler profiles of Erlangen CRM. Even with this simplification, type statements alone are 302,149,587 or 37% of the total. The number of types is 238. We counted statements per type with this query and present some of the top types:

```
select ?t (count(*) as ?c) {?o a ?t} group by ?t
```

Class	**Statements**
owl:Thing	36485904
E1_CRM_Entity□	36485903
E77_Persistent_Item□	17408450
E70_Thing□	17339714
E71_Man-Made_Thing□	17216212
E72_Legal_Object□	17192518
E28_Conceptual_Object□	14776488
E90_Symbolic_Object□	14629292
E2_Temporal_Entity□	11924877
E4_Period□	11924877
E5_Event□	11922986
E7_Activity□	11796470
E63_Beginning_of_Existence□	6377421
E11_Modification□	6296015
E12_Production□	6295825
rso:FC70_Thing	2051797
skos:Concept	415509

Comments (look at the class hierarchy as well): we have 415k terms (skos:Concept) and 2M museum objects (FC70_Thing). These objects have 6.3M E12_Production records, which are repeated as the super-class E11_Modification; there are a few hundred Repairs mapped to E11, over and above the E12 number. E12 is also repeated as E63_Beginning_of_Existence; which has additional 100k records of Birth and Formation for the Person-Institution thesaurus. Another 5.4M E7_Activity records stand for Acquisition, Discovery, exhibition, etc. Each E7 is repeated as E5_Event, which is repeated as E4_Period (with an extra 19k historic Periods) and E2_Temporal_Entity; etc.

A lot of the higher-level classes are too abstract to be useful for querying (e.g. E1_CRM_Entity, E70_Thing, E77_Persistent_Item, E72_Legal_Object. But OWLIM materializes all inferences and unfortunately doesn't offer options for controlling which ones to materialize.

2.8 Complexity: Properties

(Note: the analysis below is based on a slightly older version of the repository with 806M statements instead of 917M statements. But the percentages and conclusions are approximately the same.)

We have a total of 339 properties. We analyzed the statement distribution per property with this query:

```
select ?p (count(*) as ?c) {?s ?p ?o} group by ?p
```

Properties	**Statements**	**Percent**
rdf:type	302149587	37.50%
Objects: CRM, rdfs:label	365430152	45.35%
Extensions: BMO, RSO	35903831	4.46%
FRs (70M=9%) and sub-FRs (26M=3%)	96526377	11.98%
Thesauri: BIBO, DC, DCT, FOAF, SKOS, QUDT, VAEM	5715250	0.71%
Ontology: RDF, RDFS, OWL	4159	0.00%
Total	**805729356**	**100.00%**
CRM inverses	149465596	18.55%

- **Type** statements take a significant proportion, analyzed in the previous section.
- Statements related to O**bjects** are the majority (365M or 45% of total). Chief amongst them are P3_has_note (10.10% of the Object statements), P2_has_type (6.49%), P12_occurred_in_the_presence_of (3.29%)
  - rdfs:label is also significant (13.9M or 3.81% of the Object statements). We estimate that 5M of rdfs:label statements are due to Thesauri and should be moved from row Objects to row Thesauri.
  - A lot of the CRM properties have inverses (79 properties in our system). They are useful when writing rules and queries, but create a significant number of duplicate statements (18.6% of total: included in row Objects, and shown separately on the last row)
- **Extensions** are sub-properties of CRM, following the CRM extensibility guidelines. CRM itself uses sub-properties extensively. The maximum depth of the property hierarchy is 4, e.g.: P12_occurred_in_the_presence_of> P11_had_participant> P14_carried_out_by> P22_transferred_title_to.

2.9 Comparison to Other Repositories

Below is a comparison of the RS CRM repository to some repositories hosted by Ontotext and PSNC and provided as SPARQL public services. In each cell we show the absolute number (in Millions, except for Expansion and Density) and the ratio compared to RS CRM. **Expansion**=Total statements/Explicit statements shows the intensity of inference. **Density**=Statements/Nodes shows the relative density of the graph. **Objects** is not defined for the last two repositories, since they cover broad domains and the objects are too heterogeneous.

Repo	Objects		Expl.st.		Ex.st/obj		Total st.		Expans.		Nodes		Density		Reasoning
CRM	2.0	1	195	1	90	1	916	1	4.7	1	54	1	17.0	1	rdfs+tr+FR
PSNC	3.1	1.5	234	1.2	75	0.83	535	0.58	2.3	0.49	60	1.1	8.9	0.52	rdfs-subCl
EDM	20.3	9.8	998	5.1	50	0.56	3798	4.1	3.8	0.8	266	4.9	14.3	0.84	owl-horst
FF			1673	8.6			3211	3.5	1.9	0.4	456	8.4	7.0	0.41	owl-horst
LLD			6706	34			10192	11	1.5	0.3	1554	29	6.6	0.38	rdfs+trans

- RS CRM: http://test.researchspace.com:8081/sparql, the subject of this paper
- PSNC: Polish Digital Library: http://dl.psnc.pl, national aggregation using FRBRoo and CRM. Subclass inference is disabled to avoid a proliferation of type statements (see section 2.7). See section 4.2 for more details about his repository.
- EDM: http://europeana.ontotext.com: Europeana data, snapshot of 14.9.2012.
- FF: http://www.factforge.net [6]: an RDF warehouse and Reason-able View including 10 of the most important LOD datasets of general interest: FreeBase, DBpedia, GeoNames, MusicBrainz, etc. FactForge reasoning is described in [7]
- LLD: http://linkedlifedata.com: a semantic data integration platform for the bio-medical domain

Observations: The RS CRM repository is of moderate size compared to others (but is expected to grow as more partners join RS). CRM expresses objects in considerably more detail than all other repositories, even EDM. This can be seen in both ratios **Ex.st/obj** (explicit statements per object) and **Density** (total statements per node).

3 Performance

3.1 Performance of SPARQL Implementation

FRs can be implemented by composing straight SPARQL queries. For example, the query for FR92i_created_by (sec. 2.3) can be defined like this using SPARQL 1.1 Property Paths [8], and you can try it at the RS CRM endpoint[10]:

```
select ?obj $act {
 ?obj a rso:FC70_Thing;
  (crm:P46_is_composed_of|crm:P106_is_composed_of|crm:P148_has_component|crm:P128_carries)*/
  (crm:P31i_was_modified_by|crm:P94i_was_created_by) / crm:P9_consists_of* /
   crm:P14_carried_out_by / crm:P107i_is_current_or_former_member_of*
 $act
} limit 20
```

The first few objects returned are Rembrandt paintings from the RKD dataset. $act is bound to rkd-artist:Rembrandt, and also to groups that he belongs to: profession/draughtsman, profession/printmaker, nationality:Dutch (conversely, the user can search by such groups).

In the RS system, $act is bound to an input variable and ?obj is the output variable:

[10] http://test.researchspace.org:8081/sparql (ask the authors for login)

```
select distinct ?obj {
 ?obj a rso:FC70_Thing;
   (crm:P46_is_composed_of|crm:P106_is_composed_of|crm:P148_has_component|crm:P128_carries)*/
   (crm:P31i_was_modified_by|crm:P94i_was_created_by) / crm:P9_consists_of* /
    crm:P14_carried_out_by / crm:P107i_is_current_or_former_member_of*
 rkd-artist:Rembrandt
} limit 20
```

The endpoint takes over 15 minutes to answer the query. If you add more clauses, the performance is even worse. The query can be optimized a bit by using intermediate variables instead of property paths, but the performance is still untenable.

3.2 Performance of FR Implementation

The same query, using FR92i_created_by as defined in sec. 2.3, is trivial and has **sub-second response time**:

```
select distinct ?obj {?obj rso:FR92i_created_by rkd-artist:Rembrandt} limit 500
```

Currently RS imposes a limit of 500 results due to browser memory limitations of the used faceting system (Exhibit 2), but even the full set of 1418 objects is returned within a second.

Now let's add some complexity: let's find **drawings** by Rembrandt that are about **mammals**. We first need to find the corresponding thesaurus terms, e.g.

```
select * {?s rdfs:label "drawing"}
select * {?s rdfs:label "mammal"}
```

The query uses another FR from the list in sec. 2.1: FR2_has_type (which is used to relate to any E55_Type term, no matter whether it relates to the **isness** or **aboutness** of the object):

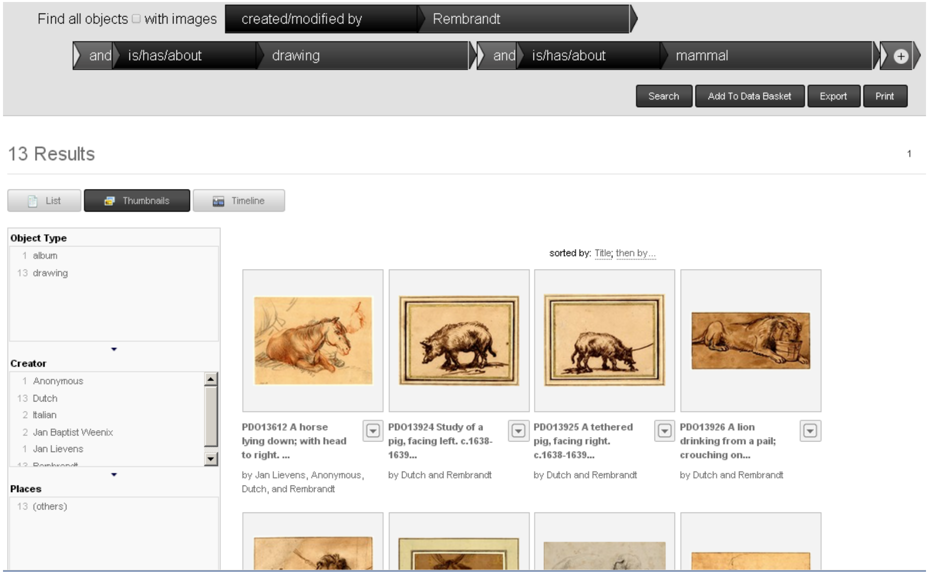
```
select distinct ?obj {
 ?obj rso:FR92i_created_by rkd-artist:Rembrandt;
     rso:FR2_has_type thes:x6544, thes:x12965
} limit 500
```

The query takes less than a second and returns 13 objects. None of them has subject "mammals" per se: they are about horses, pigs, lions, camels and an elephant (see next screen-shot). But the corresponding FR is defined as transitiveOver skos:broader, so it navigates the term hierarchy.

Materializing the FR triples adds 12% to the repository size (see sec. 2.8), which has negligible slow-down on basic querying speed. As shown in sec. 2.9, OWLIM has been used successfully on much bigger repositories, so this extra size is not a concern.

3.3 RS Semantic Search

RS uses the above to implement Semantic Search with controlled vocabularies and faceting. The user enters terms using auto-completion, RS restricts to FRs applicable to the specific term (e.g. created/modified is applicable to Agents, whereas is/has/about is applicable to concepts such as Object type, Subject, etc) and constructs a "search sentence". Here is a screen shot; you can view a video[11] of RS search in action, or ask the authors for a demo.



Find all objects □ with images | created/modified by | Rembrandt

and | is/has/about | drawing | and | is/has/about | mammal

13 Results

Object Type

1 album
13 drawing

Creator

1 Anonymous
13 Dutch
2 Italian
2 Jan Baptist Weenix
1 Jan Lievens
13 Rembrandt

Places

13 (others)

sorted by: Title; then by...

PDO13612 A horse lying down; with head to right. ...
by Jan Lievens, Anonymous, Dutch, and Rembrandt

PDO13924 Study of a pig, facing left. c.1638-1639...
by Dutch and Rembrandt

PDO13925 A tethered pig, facing right. c.1638-1639...
by Dutch and Rembrandt

PDO13926 A lion drinking from a pail; crouching on...
by Dutch and Rembrandt

Note: the **Creator** facet is populated from FR92i_created_by, which includes not only individual creators but also groups they belong to. In this case "Dutch" is the Nationality of Rembrandt.

This search uses the query defined in the previous section. The search takes significantly longer than the query alone (4.5 seconds) because after obtaining up to 500 objects, it executes several more queries to fetch their display fields, facets, and images. Subsequent restrictions using the facets are much faster (sub-second response).

4 Conclusion

4.1 Summary

We presented performance results for the RS implementation [1] of FR Search as defined in [2,3]. This search works over a significant CH dataset (almost 1B statements), using a complex ontology (CIDOC CRM). Using a semantic repository is appropriate for this dataset because of its complexity, graph-oriented nature, diversity

[11] http://www.youtube.com/watch?v=HCnwgq6ebAs

of relations, and complexity of queries that users are interested in. This is an exciting demonstration of large-scale reasoning with real-world data:

- The well-structured nature of the data allows for expressive reasoning. The inferred knowledge makes good sense when reviewed by domain experts; unlike other combinations of RDF data gathered "from the wild" that often generate strange/ faulty results.
- This is one of the first examples of such expressive reasoning with large datasets. Previous examples work with 5-10M statements, and often use synthetic data
- Reasoning adds real value: it would be very hard to service the same complex queries without inference

4.2 Related Work

The RS repository is one of the largest CH datasets loaded in an RDF repository and provides valuable implementation experience. Some other large CH repositories include:

- The Europeana EDM repository (hosted by Ontotext) is bigger (see sec. 2.9), but is much less structured. Since most objects were converted from ESE, they include mostly literals: no controlled URIs and no links.
- CLAROS[12] (Classical Arts Research Online Services): in 2009 [9] reports 10M triples loaded in a Jena TDB triple-store. This has expanded, implementing offline indexing extensions (MILARQ) for better performance.
- The Poznan Supercomputing and Networking Center (PSNC) has implemented several national aggregations of museum and bibliographic data based on CIDOC CRM, also using OWLIM. [10] reports 600k publications converted to CRM/ FRBRoo as part of the SYNAT project. Krzysztof Sielski reported the numbers in section 2.9 at the CRMEX 2013 workshop, see the PSNC paper in this volume

4.3 Future Work

We would like to re-implement the FRs by using a lot of standard constructions and only a few OWLIM rules:

- Standard RDFS and OWL constructs: rdfs:subPropertyOf, owl:propertyChainAxiom, owl:inverseOf
- Additional properties: ptop:transitiveOver as generalization of owl:TransitiveProperty; conjunctive property definitions that are needed for FRs, as explained in [1] sec. 3.3
- Define the FR networks in RDF data

The benefits of such reimplementation are better flexibility (OWLIM rules are not flexible, see end of sec 2.2) and better portability to other repositories.

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

We are exploring the opportunity to create a CH Benchmark using the above data and FRs under the auspices of the Linked Data Benchmarking Council (LDBC)[13] project, so that other vendors can implement the same reasoning and compare the performance of their implementations. LDBC seeks to empower users of semantic technologies by establishing significant and objective benchmarks that address real-world data and user needs.

4.4 Acknowledgements

This work is supported by the ResearchSpace project, executed by the British Museum and funded by the Andrew W. Mellon Foundation. We are grateful to Dominic Oldman, ResearchSpace principal investigator, for his excellent continuing collaboration. Dominic created the RS Semantic Search movie cited above.

We thank the anonymous reviewers for their constructive and specific comments.

5 References

1. Vladimir Alexiev: Implementing CIDOC CRM Search Based on Fundamental Relations and OWLIM Rules. Semantic Digital Archives workshop (SDA 2012), part of Theory and Practice of Digital Libraries conference (TPDL 2012). September 2012, Paphos, Cyprus. http://ceur-ws.org/Vol-912
2. Katerina Tzompanaki, Martin Doerr: A New Framework for Querying Semantic Networks. ICS-FORTH Technical Report TR-419, May 2011
3. Katerina Tzompanaki, Martin Doerr: Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories, ICS-FORTH Technical Report TR-429, April 2012, http://www.cidoc-crm.org/docs/TechnicalReport429_April2012.pdf
4. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov, OWLIM: A family of scalable semantic repositories, Semantic Web Journal, Volume 2, Number 1, 2011.
5. Barry Bishop, Spas Bojanov. Implementing OWL 2 RL and OWL 2 QL rule-sets for OWLIM. OWL Experiences and Directions workshop (OWLED 2011), San Francisco, USA, June 5-6, 2011, CEUR-WS.org, ISSN 1613-0073
6. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov. FactForge: A fast track to the web of data. Semantic Web Journal, V.2, N.2, 2011.
7. Barry Bishop, Atanas Kiryakov, Zdravko Tashev, Mariana Damova, Kiril Simov. OWLIM Reasoning over FactForge. Proceedings of OWL Reasoner Evaluation Workshop (ORE'2012), collocated with IJCAR 2012, Manchester, UK
8. SPARQL Property Paths, http://www.w3.org/TR/sparql11-property-paths
9. David Shotton, The Future of the Past: Using CIDOC CRM for CLAROS. Semantic Web and CIDOC CRM Workshop, co-located with ISWC 2009, Washington DC, http://www.semuse.org/index.php?title=Semantic_Web_and_CIDOC_CRM_Workshop
10. Cezary Mazurek, Krzysztof Sielski, Maciej Stroiński, Justyna Walkowska, Marcin Werla, Jan Węglarz. Transforming a Flat Metadata Schema to a Semantic Web Ontology: The Polish Digital Libraries Federation and CIDOC CRM Case Study. Studies in Computational Intelligence Volume 390, 2012, pp 153-177

[13] http://www.ldbc.eu/

Simple visualization of structures of interrelated concepts in the FRBRoo ontology

Krzysztof Sielski, Marcin Werla

Poznań Supercomputing and Networking Center,
ul. Noskowskiego 12/14, 61-704 Poznań, Poland
{sielski,mwerla}@man.poznan.pl

Keywords: Semantic Web, ontology, OWL, RDF, FRBRoo, RDF Unit, knowledge base

The Knowledge Base which was created by Poznań Supercomputing and Networking Center (PSNC) as a part the SYNAT[1] project integrates information from distributed heterogeneous sources such as digital libraries, digital museums, scientific and technical information systems. The gathered knowledge is stored in an RDF semantic database and is represented in FRBRoo ontology with some custom extensions, which had to be introduced in order to represent all the information without any semantic loss.

As of the beginning of 2013, the Knowledge Base contained information from over 3,100,000 metadata records, which were originally encoded in various schemas: PLMET (data obtained from Polish Digital Libraries Federation), MARC 21 XML (from union catalog of Polish research libraries NUKAT), MONA (from the National Museum in Warsaw) or CDWA LITE (from the National Museum in Krakow). These records were converted to FRBRoo ontology using jMet2Ont[1] tool. Some auxiliary data sources such as VIAF, Geonames, KABA Subject Headings and Lexvo have been used to enrich the records with detailed information. Currently, the number of RDF triples building the Knowledge Base is 536M, which includes 235M explicit and 301M implicit triples. The implicit triples have been added by the inference engine with our custom rule set, which is a subset of OWL 2 RL/RDF entailment rules.

Unlike traditional relational databases, data represented as triples does not have a precise schema with strict constraints. Instead, OWL ontologies describe the structure of concepts and relations between them. As FRBRoo is a complex ontology with many classes, this model is often converted to a simpler one when presented to an user. The contents of the Knowledge Base can be explored in a couple different ways:

- a raw SPARQL endpoint, which is aimed at expert users who know the ontology very well and have precisely defined goals;

[1] SYNAT project, financed by Polish National Center for Research and Development (grant number: SP/I/1/77065/10), is aimed to conduct a research task titled “Creation of universal, open, repository platform for hosting and communication of networked resources of knowledge for science, education and open society of knowledge”

- a full text search application, which searches for keywords provided by user in RDF literals from the triplestore and uses the Query Processing Module (QPM) which maps on-the-fly information represented in the FRBRoo ontology to a simplified model, consisting of the following concepts: works, items, persons, places, legal bodies, and subjects;
- a geographical search application, which allows user to select an area on a map to find all objects connected with places contained in that area (e.g. all publications whose subject is a particular city);
- an application to explore semantic database with dynamically fetched portions of data describing particular object from the triplestore, which are presented as interrelated FRBRoo concepts in a legible way understandable by non-experts.

The last named application was built as a proof of concept of RDF Unit[2]. RDF Units are graphs which consist of several ontology objects of different classes that are needed to provide all the essential information about a certain resource. For example, an RDF Unit for a particular instance of Publication Expression from the Knowledge Base would include objects representing its Title, Publication Event and Place of Publishing, but not geographical coordinates of that place. RDF Units are dynamically constructed based on the metaproperties of ontology relations and actual data in the triplestore.

Such graphs are transformed into a tree structure, in which the examined resource becomes a root. Then, the obtained RDF Unit tree is prepared for presentation by replacing names of predicates with more user friendly labels and by flattening some long predicate paths to a single dummy edge in order to provide information in a straightforward way. This transformations are represented as a set of rules which take into consideration a predicate and classes of a subject and an object. Examples of such rules include (here `[?]` stands for any class):

- `[E21_Person] P100_i_died_in [E69_Death] P4_has_time_span [E52_Time-Span] P1_is_identified_by [?]`→ *date of death*
- `[F18_Serial_Work] P148_has_component [F14_Individual_Work]` → *series element*
- `[?] P9_consists_of [F28a_Contribution] P14_carried_out_by [?]` → *contributor*
- `[?] P9_consists_of [F28a_Contribution] P2_has_type [?]` → *in the role of*

Figure 1 presents a result of mapping one record in MARC 21 XML schema to FRBRoo. It is a graph of 47 connected FRBRoo objects represented by 108 RDF triples. Figure 2 presents a view in our application that represents an RDF Unit of an Individual Work resource which was created in mentioned mapping. This unit contains all the information from source record except for author's and contributors' dates of life, which can be examined in those resources' view.

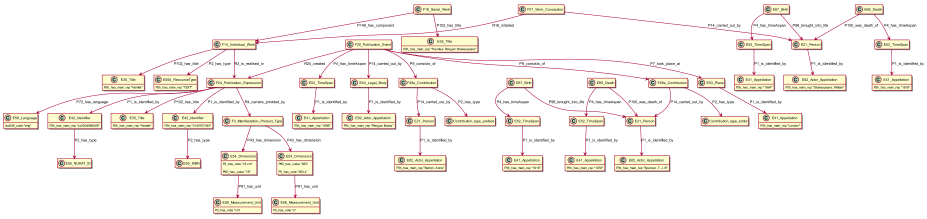
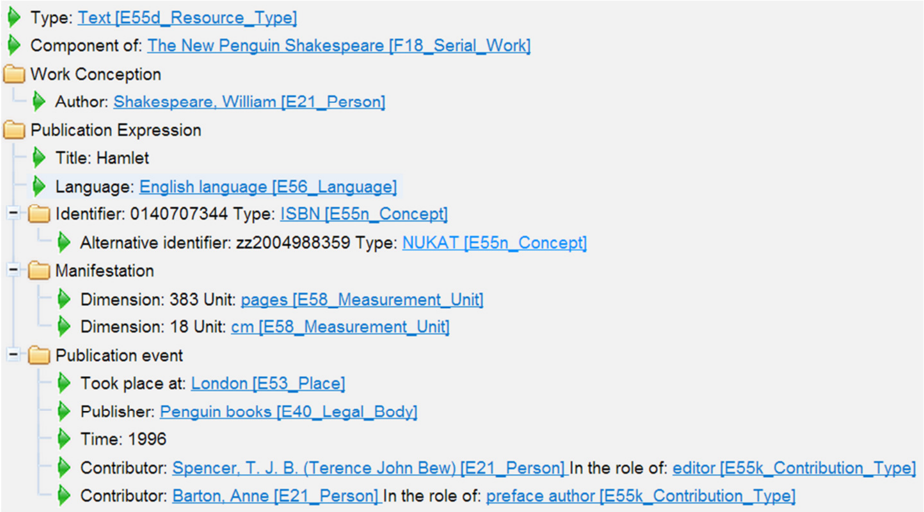


Fig. 1 A result of mapping a single metadata record from MARC XML to FRBRoo represented as a graph. An image in high resolution can be viewed at http://bit.ly/frbroo_ham



Type: Text [E55d_Resource_Type]

Component of: The New Penguin Shakespeare [F18_Serial_Work]

Work Conception

- Author: Shakespeare, William [E21_Person]

Publication Expression

- Title: Hamlet
- Language: English language [E56_Language]
- Identifier: 0140707344 Type: ISBN [E55n_Concept]
  - Alternative identifier: zz2004988359 Type: NUKAT [E55n_Concept]
- Manifestation
  - Dimension: 383 Unit: pages [E58_Measurement_Unit]
  - Dimension: 18 Unit: cm [E58_Measurement_Unit]
- Publication event
  - Took place at: London [E53_Place]
  - Publisher: Penguin books [E40_Legal_Body]
  - Time: 1996
  - Contributor: Spencer, T. J. B. (Terence John Bew) [E21_Person] In the role of: editor [E55k_Contribution_Type]
  - Contributor: Barton, Anne [E21_Person] In the role of: preface author [E55k_Contribution_Type]

Fig. 2 A representation of F14_Individual_Work object (*Hamlet* by Shakespeare) from Fig 1 converted to a simplified tree compatible with FRBRoo ontology for presentation

The described Knowledge Base browser application was prepared for dynamic viewing of FRBRoo data from the triplestore, but this approach is generic and should work for another ontologies as well. It uses no predefined SPARQL queries and is based only on a relatively small configuration: a set of graph path flattening rules for presentation and a set of single metaproperty for each ontology predicate which is used to build RDF Units.

References

1. Walkowska, J., Werla, M.. (2012). Advanced Automatic Mapping from Flat or Hierarchical Metadata Schemas to a Semantic Web Ontology. TPDL 2012. Lecture Notes in Computer Science, vol. 7489, pp. 260-272.
2. Sielski, K., Walkowska, J., Werla, M.. (2013). Methodology for Dynamic Extraction of Highly Relevant Information Describing Particular Object from Semantic Web Knowledge Base. TPDL 2013. Lecture Notes in Computer Science, vol. 8092, pp. 260-271.

CRMEX Position Paper: mapping patterns for CIDOC CRM

Douglas Tudhope, Ceri Binding

Different mappings and different implementation primitives can potentially pose significant problems for semantic interoperability, as discussed in our workshop paper. It would be useful if we could provide more examples and guidelines. Examples of some possible mapping issues from our experience include

- Should an E57 Material (e.g. *gold)* be mapped as a property of an E11 Modification event or as a property of an E22 Man-Made Object?
- Should a method of manufacture (e.g. *hammered*) be mapped as an E55 Type of an E12 Production event or as an Appellation of an E29 Design or Procedure?
- Should E22 Man-Made Objects be directly identified by an E42 Identifier or should the connection be made via a record that has an Identifier?
- All CRM classes can be assigned types (used for domain terminology) – any guidelines for default choices?
- When is it appropriate to create an assignment event when assigning an attribute to an object? Essentially this depends whether the decision to assign an attribute is considered worthy to record. Again this can result in different mapping expressions depending on the judgement.
- How to deal with short cuts in a consistent way and link a shortcut with its underlying path?

The potential to employ reasoning over the CRM graph is one of the reasons for semantic integration. Nonetheless in our view, a multiplicity of approaches for similar data will pose unnecessary problems for implementation in the medium term. Specific rules will probably be required, which raises difficulties for generalising and introducing a new alternative mapping. A pragmatic approach is to combine developments in reasoning with efforts at consensus on patterns for CRM mappings and guidelines. This could involve patterns for particular domains and also general patterns for common situations.

Issues

If different implementations of the CRM follow different low level implementation specifications or employ different mappings for the same underlying semantics then this raises barriers for semantic interoperability.

Working from established RDF patterns guarantees the semantic interoperability of the resultant data and also that the syntactical implementation details are handled consistently. It is also more friendly to non-specialists and can make it easier to express datasets via CIDOC CRM.

- Agreement on implementation details (e.g. primitives, namespace, definitive URIs)?
- Agreement on mapping patterns and guidelines?
- Desirability of expressing the end-purpose or use cases of a mapping exercise?
- Provision of appropriate registries of mapping patterns?
- Provision of core metadata for mapping patterns?