Online event-based conservation documentation
A case study from the IIC website

Athanasios Velios

21st November 2013

Introduction [heading]

The term *conservation documentation* is typically used to describe records about objects or monuments. These records are used to a) document the structure and condition of objects and monuments, b) record previous and current conservation work alongside its evaluation and c) justify the choice of conservation materials and techniques for an intervention. Although the value of documentation is emphasised in conservation training and it is a task undertaken regularly in the profession, in depth discussions by conservators on documentation practices are rare. The recent redevelopment of the IIC website, which is a large online conservation resource, is a good opportunity to engage in such a discussion and this paper will hopefully function as a starting point.

In most conservation records the common characteristic of documentation is that they are produced to accompany an object. Retrieving a record is done as part of investigating the object and rarely is a record considered independently. However, with the increasing popularity of the internet as a medium for sharing data, there is great scope in considering documentation records as separate entities which can offer valuable information in themselves, even when the focus is not on the corresponding objects. Sharing documentation records is valuable because the larger the number of available records, the better they represent conservation activity and therefore the more significant they become for statistical analysis[1]. Some examples of the value of documentation data are included in the next section (Previous attempts [heading]).

However, querying conservation records from diverse sources is currently impossible because they do not conform to a common structure. As discussed in Structured documents: schemas [subheading] it appears that adopting a "common structure" for conservation documentation is utopic and unreasonable. It is utopic because it is impossible to force all conservators who publish records online to adopt one documentation format. It is unreasonable because each object often requires special treatment and its documentation record does not necessarily conform to existing documentation structures. This is a generic problem in many professional fields and it has been considered as part of the proposal for the *Semantic Web*. The adopted solution indicates that by making documentation structures abstract, it is possible to adopt a common documentation framework. Later on I will refer to the *Resource Description Framework* (RDF) as a common abstract documentation framework for conservation within the overall proposal of the Semantic Web.

In addition to its traditional use, the term *conservation documentation* can describe records about conservation resources. Examples include publications and news items related to conservation which are found on conservation websites. In this document, *conservation documentation* is used with this wider meaning.

Many of the ideas discussed in this document are borrowed from the fields of *computer science* and *knowledge organisation*, which I believe are under-

[1] One may use the term *Conservation Big Data* to describe this.

represented but extremely valuable in conservation documentation. Of particular use is the work of the International Committee for Documentation of the International Council of Museums (ICOM-CIDOC)[2], who have undertaken pioneering work to provide a standard documentation model in cultural heritage, including conservation (see CIDOC-CRM, AAT and RDF [heading]). Likewise the Getty vocabularies[3] are particularly important resources which can serve conservation documentation well.

Previous attempts [heading]

Free text [subheading]

A good account of the development of documentation in conservation can be found in Ravenberg [2012]. Ravenberg reviews a range of projects from the past 30 years featuring innovative (at the time) projects about conservation documentation. She compares these with current documentation practices used in museums and she concludes that documentation in conservation today is primarily done using free text in software which resembles the paper format (i.e. text processors). It seems that conservation has adopted computer technology in documentation to replicate the old paper forms which were standard practice since the establishment of the profession and which mostly relied on free text. Adopting free text in computerised conservation records has an important benefit: familiarity. A conservator who is used to filling in conservation documentation forms on paper, can easily undertake the same process in front of a computer. However, the price to pay for familiarity is high, as discussed next.

In the 1980s and 1990s conservators considered that the concept of information retrieval typically meant time-consuming and repetitive process of going through piles of sheets of paper. Because the primary purpose of looking up records was to find out about previous treatments of an object, the solution of the free-text paper form served this purpose well. It takes the conservator a couple of minutes to find the record for a specific object in a well-organised archive and 5-10 minutes to read through it and get a good idea of what previous treatments have taken place. However, as mentioned in the introduction, per-object information retrieval is not the only way to use conservation records. There is value in searching for information **across** collections. For example, if it were possible to classify conservation techniques as interventive or non-interventive, by querying a representative volume of online conservation data we could discover the percentage of conservation activity which has been interventive. Then we could correlate this result to the ownership of objects and test the assumption that objects from private collections are "more heavily" conserved than objects in public museum collections thus informing the discussion around conservation ethics. Another example would be when evaluating the success of specific materials or techniques in fulfilling their intended use in order to select the most

[2] http://network.icom.museum/cidoc/

[3] http://www.getty.edu/research/tools/vocabularies/index.html

effective one for future work. In paper free-text archives the value of documentation records is limited because information retrieval across the collection is impractical. It would involve a conservator having to read every document which could take many days of work. This work is rarely done and the result is that documentation records are not being used comparatively.

Conservation records currently published on websites (or held in databases) have many similarities with the paper equivalents in that they mostly rely on free text. Retrieval of the record by the computer is much faster but reading through the text of the record is again time-consuming. This is a problem in many professional fields and the objective of the field of *information extraction* is to replace time-consuming reading by humans with a digital system which will read and analyse text, isolate useful information and return this as the essence of the meaning of the text without human intervention. Studies in this field have offered promising results in sectors with standardised onomatology (such as the bio-sciences), but they are still far from replacing human understanding. Earlier studies (e.g. Soderland, 1999) showed how results from software processing free text have limited applications and they could hardly be compared to results from human processing. In more recent reviews (e.g.[Chang et al., 2006, § 5.1]) the conclusions are similar. These studies indicate that *structured* or *semi-structured* documents are preferable for automatic text analysis than free text documents (figure 1).

A structured document is a document which follows a predefined template where each line/field is clearly documented with metadata alongside the possible values it may take. Typically a structured document accepts monolectic (or rather single concept) data for each field. For example, if we define a field in our document record for the type of a Greek vase, we need to document the field (i.e. the fact that the morphological type should be described there) and also document the possible vase types, so that the reader is clear on the differences between, say, an *amphora type A* and an *amphora type B*. A semi-structured document is loosely defined as a document where some information is stored in a pre-defined structure, and other information is within free text fields.

Structured documents: schemas [subheading]

As mentioned in the previous section, research indicates that data retrieval from structured documents returns better results than data retrieval from free text. Conservation has adopted this principle but only with minimal effect and mostly empirically. Structured documents are rarely used, but semi-structured documents are more frequent i.e. free text is inserted within headings which outline a basic document structure. For example an object record may include a heading such as *previous repairs*, but the description of previous repairs is done using free text. This is an insufficient step towards effective machine-based retrieval because it still relies on free text and therefore requires a human to read the text and digest the information.

These documents are produced by replicating a template and it is not unusual to have slight modifications to the template to better accommodate the

requirements for describing a specific object. For example, the survey form of the Saint Catherine Library Conservation Project evolved during the project as documented by Pickwoad (2004, p. 35) and it introduces an important problem in machine-based searching: that of consistency. If records from a collection are based on multiple templates, there is an added complication in retrieving results in a digital system because it has to be programmed separately for every template used. The problem becomes more complex if there are no user instructions accompanying each template and this is not unusual for digital forms let alone for paper based forms. For digital records, the solution to this problem of consistency is given by the adoption of a so-called *schema*.

A schema formalises the form template into an explicit set of conditions which allows consistent digital records to be produced. The schema defines a set of questions which the conservator is required to answer while completing the record. The schema ensures that all questions marked as obligatory are answered by reminding the user of missing answers. It validates answers and rejects the ones which are out of context. The schema defines the kind of data expected in an answer (e.g. a *number* is required for a dimension) and will notify the user if anything other than the expected type of data is inserted. Schemas can be defined using a variety of tools. Relational database tools became mainstream in the early 1990's with a wealth of user-friendly software and are relatively easy to create schemas with. More recently the *Extensible Markup Language* (XML) documents have increased in popularity and schemas for these can be built with a range of languages including the W3CSchema[4] and RelaxNG[5]. For example, the Saint Catherine survey made the transition from paper to digital by implementing a schema for XML documents[Velios and Pickwoad, 2009].

Although, the adoption of a schema is a big step towards machine-based searching it has a major drawback: it only reflects the requirements at a local level. A schema produced for the requirements of a specific project rarely matches the requirements of another. This is true for the various implementations of schemas for semi-structured data in museums. Standards such as Spectrum 4.0 [Dawson and Hillhouse, 2011] are widely adopted but they do not require specific schemas. Therefore, data produced and organised following one schema may look very different to data produced following another. This means that the schema helps to produce consistent digital data within an organisation, but the problem of consistency is now moved to a different level because records of one organisation do not match records of another. Traditionally this problem is solved by standardisation, i.e. a group of leading organisations recommends a standard schema which is universally adopted. This has never happened for conservation and in the past, proposals for "best" schemas have not been widely adopted and they would be difficult to enforce. This is true for information held on conservation websites which have been developed independently and do not conform to a common structure[6].

[4] http://www.w3.org/XML/Schema

[5] http://relaxng.org/

[6] Various commercial systems which are likely to be used in large organisation may have attempted to take over the market and as such enforce some consistency, but in many cases

In the field of *knowledge organisation* consistency has been achieved at a higher level: by building a generally accepted *concept thesaurus*. A concept thesaurus is a set of concepts which can be used in any schema. Because the concepts of a domain (in our case conservation) are globally accepted (even if the terms used to describe them are not), it is possible that there could be universal agreement in an inclusive concept thesaurus, thus achieving some conformity in schemas.

Controlled vocabularies and thesauri [subheading]

On conservation websites, records are typically grouped using classification terms. The terms typically form an index menu where a user can click to retrieve relevant documents. These terms could be considered as controlled vocabularies for conservation classification but no systematic work on the formation of these vocabularies has been done and these indexes have been produced on an ad-hoc basis. Three examples of such indexes are shown in table 1.

Let us consider the term *silver* for example. We anticipate that all users visiting these websites accept the same meaning for this term: *documentation about the conservation of silverware*. Accepting a different meaning (for example *silver* as in *silver paint*) would reduce the value of the index, if not make it completely useless. Therefore the successfulness of an index relies on contextual understanding by users, which is not guaranteed.

Another interesting observation to make is that users are able to understand that although *silver* and *ethnographic material* are two different categories, this does not mean that silver objects cannot be found in an ethnographic collection. Similarly with *furniture* and *decorative surfaces*, they appear to be separated, but furniture can consist of decorative surfaces. Users are able to approach an index intelligently to eliminate such trivial problems. A digital system, however, would not be as successful in using the same index because it does not have the understanding of the context. Concept thesauri could be used to make such indexes accessible to digital systems without human intervention.

The terms in the above indexes can be grouped as follows:

- conservation of *techniques*, which typically describe the making of objects (e.g. *gilding conservation*),
- conservation of *attributes*, which focus on important properties of the object (e.g. *polychrome sculpture conservation*),
- conservation of *materials*, which focus on the main material of the object (e.g. *ceramic conservation* could include both thrown ceramic vases and modelled ceramic sculpture),
- conservation of *types of objects*, which focus on the form or function of the object (e.g. *building conservation*).

these largely ignore any significant level of detail in conservation documentation as Ravenberg [2012]describes.

AIC	ICON	AICCM
Architecture[a]	Archaeological	Book and paper
Books	Books	Digital media
Glass and Ceramics	Carpets and rugs	Disaster preparedness and risk management
Documents and Works of Art on Paper	Ceramics and glass	Objects
Furniture	Clocks	Objects → Archaeological materials
Metal Objects	Costume and textiles	Objects → Plastics
Paintings	Furniture	Objects → Machinery and scientific equipment
Photographs	Oil paintings	Objects → Wood and furniture
Textiles	Photographic material	Objects → Ethnographic collections
Matting and Framing	Documents and archives	Objects → Ceramics and glass
Disaster Response and Recovery	Prints, drawings and watercolours	Objects → Metals
Health and Safety	Silver and plate	Objects → Handling objects
	Decorative schemes and surfaces	Photographs, film and audiovisual material
	Ethnographic objects	Paintings and frames
		Storage and the environment
		Textiles
		Time capsules

Table 1: Example of ad-hoc index for conservation documents

[a]The term is used instead of the term *building*.

At the same time many of these indexes include terms which are not linked with the object being treated:

- conservation *functions*, which are generic activities undertaken within the conservation profession (e.g. *pest control*),
- conservation *disciplines*, which indicate conservation sub-domains (e.g. *preventive conservation*).

Classifying a conservation record requires the use of more than one term - highlighting different aspects of the object. Let us attempt to group the terms of table 1 based on the list of aspects from our grouping. A possible result of this grouping is shown in table 2. There are a few comments emerging from this process:

1. The terms used in the example indexes are often compound terms or terms with double meaning. This is bad practice for machine readability since it is difficult for a digital system to identify the context within which the term is used in order to choose the right meaning for it. For example *wood and furniture* is used as one term but it applies to both *wood* as material and *furniture* as the type of object. Programming a digital system to understand whether a record is about wood in general or furniture could be a challenging process. Another example: *oil paintings* are *paintings* and should be classified under that type of object. However, the purpose for specifying *oil paintings* instead of *egg-tempera paintings* or simply *paintings* is to give an indication of the material and as such one could argue that the term should also go under materials. Ideally, for machine searching, each term in the index should only describe a single conservation concept.

2. Most of the terms used fall under two main categories, *materials* and *type* of objects. This is not a surprise given that conservation departments have been set-up according to material and type in museums and educational institutions (e.g. sculpture conservation, metal conservation). This is particularly interesting when compared with the terms *archaeological* and *ethnographic* conservation which perhaps are rarely found as museum department names but do exist in educational departments. These terms have been used in the past to highlight a set of considerations about the objects treated (for example religious considerations about liturgical ethnographic objects) and they should not be used to characterise the typology of objects involved.

3. In many instances, the same terms have the same meaning across the three examples. Ceramics conservation in one index almost certainly has the same meaning in the others. Because the meaning of these terms is the same we can envisage a more widely adopted concept thesaurus to cover all indexes.

Aspect	Terms
Technique	Silver and plate
Attribute	Decorative schemes and surfaces
Material	Glass
	Ceramic
	Metal
	Textile
	Oil paintings
	Silver and plate
	Photographs, film and audiovisual material
	Wood and furniture
	Plastics
Type	Architecture
	Books
	Documents
	Furniture
	Paintings
	Photographs
	Carpets and rugs
	Clocks
	Costume
	Oil paintings
	Photographs, film and audiovisual material
	Time capsules
	Machinery and scientific equipment
	Paintings and frames
Function	Handling objects
Discipline	Health and Safety
	Archaeological
	Ethnographic
	Disaster Response and Recovery
	Disaster preparedness and risk management
	Storage and the environment

Table 2: Grouped terms according to highlighted aspect.

4. Although these indexes have been produced as an entry point to documentation and are therefore fixed, in many cases, when filing a new document, the requirement for a new term emerges. For example, a list of documents on *metal conservation* may include a large number of documents about *iron conservation.* A new term, *iron conservation,* can be introduced as a child term of *metal conservation* making the index more detailed. There is no limit to the number of terms which can be added, but experience has shown that in some cases indexes with deeper than four or five levels become logistically difficult to manage. On conservation websites the addition of extra terms to what is meant to be an entry level index is awkward because the new terms make the index list too long. Therefore although indexes can offer detailed searching, in many cases their implementation is not fit for purpose.

There is value in adopting a universal conservation index for organising conservation content online. A recommendation for that is given in CIDOC-CRM, AAT and RDF [heading].

Object-centric versus event-centric documentation [subheading]

In the previous sections I highlighted two characteristics of conservation documentation: the use of free text and the use of ad-hoc indexes to organise free text. Another important characteristic is the fact that most documentation systems are organised around the idea of the *object* which creates problems when documenting objects with complex history. A simple example from the Saint Catherine conservation project[7] follows: a medieval binding has been altered during the long history of the book and new endleaves have been added to the existing ones. Documenting the binding in its current state (i.e. with a single set of endleaves) would mean that the fact that some endleaves are later additions would be lost. To ensure this information is recorded, the Saint Catherine's schema allows for each endleaf to be marked as *added*. However, if we consider a binding that had endleaves added to it multiple times (e.g. 100 and 300 years after it was initially bound), marking all the non-original endleaves as *added* is no longer a solution, because this does not tell us **when** an endleaf was added (i.e after 100 years or after 300 years). In this example, the object-centric approach to documenting the binding is not sufficient.

Extensive work on this has been completed by Ravenberg [2012] where the following recommendations were formalised (specifically, [Ravenberg, 2012, sec.4.1, § 4.1.1]). Let us consider that adding endleaves to the book is part of the conservation process. As standard practice, the conservator will document this activity. Typically the documentation is still done with free text but the important point here is that adding the endleaves is documented as an **activity** (i.e. *adding* endleaves), and not as an object (i.e. *added* endleaves). Because the conservation record has a date assigned to it, the activity of adding endleaves

[7] A description of the project can be found in Pickwoad [2004].

is placed in time and therefore it is distinguishable from earlier endleaf additions. It seems, therefore, that an activity-centric (or event-centric) approach is potentially more successful in recording conservation documentation than the traditional object-centric approach.

This is in agreement with recommendations made by the ICOM-CIDOC as explained in the next section (CIDOC-CRM, AAT and RDF [heading]) alongside a recommendation of how existing technologies can be used to combine the use of thesauri in an event-centric conceptual system.

CIDOC-CRM, AAT and RDF [heading]

Conceptual Reference Model [subheading]

As mentioned in a previous section (Structured documents: schemas [subheading]) it is difficult for different users to agree on a common schema for conservation documentation and the consequence is that data from different sources cannot be queried together by a digital system. Conservation is not the only field where this problem occurs. Recent reports from civil engineering and architecture such as by Zhang et al. [2011] are based on the fact that common schemas across a domain are not possible. Similar reports were published earlier in the cultural heritage sector (e.g.Baca, 2003). The main reason why there cannot be universal agreement for a single schema is because the requirements among users are different, there is no schema that will cover all requirements and it is unlikely that such a schema can be produced. The attempt to cover every requirement would make the schema logistically unusable. A user whose requirements are not successfully covered by the schema would either compromise the quality of his work or produce a different schema. At the same time a user who only requires a fraction of the schema would have a large amount of fields/questions which are not applicable[8]. Attempts to enforce conformity, through the use of either an index of terms (as mentioned in section Controlled vocabularies and thesauri [subheading]) or a schema, are unsuccessful and therefore machine searching of diverse resources is still a problem.

In section Controlled vocabularies and thesauri [subheading] I referred (in point 3) to the potential for consensus when using concepts instead of terms, which would make machine searching easier. In *computer science* this idea has been formalised with the use of *ontologies*. An ontology is a set of rules which maps the concepts and their relationships in a domain. For example, in the case of the added endleaves, the concept of the action: *adding* is linked to the concept of the object: *endleaf*, because *endleaf* is what is added during this action. *Adding* is also linked to the concept of *time* since the duration and beginning of an action can be defined. This is summarised in table 3.

[8]The Text Encoding Initiative (TEI)[2007] recognised this problem and produced a tool that will tailor the TEI schema to smaller specialised versions called Roma (Mittelbach). However, this specialisation does not necessarily mean agreement with the TEI schema as explained in the same document ([TEI Consortium, 2007, § 23.2,23.3]) which reinforces the fact that uniformity is impossible.

Generic concepts	Specific activities
action	adding
object	endleaf
time	01-10-2000

Table 3: Map of activity for adding endleaves with the corresponding concepts.

The fact that the generic concept of an action is undertaken with an object, at a certain time, forms an ontology rule for activities in conservation[9]. If the ontology rules have been tested thoroughly, it is reasonable to say that they represent reality accurately and that they can be considered robust. The technicalities of how the data is stored, i.e. the relationships of the data in the database system are irrelevant, provided that the same data can be presented to a search engine based on the pre-defined ontological relationships. In other words, if a database can be mapped to a domain ontology then a search engine will be able to retrieve data regardless of the schema or database software used. Therefore the way to circumvent the problem of schema uniformity is to develop a domain ontology.

In the cultural heritage sector there are numerous metadata schemas[Riley, 2009], but few efforts to produce an ontology. This is partly because an important project pioneered the field from early on: the CIDOC-CRM [Crofts et al., 2011]. CRM stands for *Conceptual Reference Model* and it defines a set of core concepts (*entities* or *classes*) with their relationships (*properties*). It is beyond the scope of this document to describe the various entities of the CRM. Instead an example of using the CRM to describe multiple added endleaves of a manuscript is given in figure 2. The book (CRM entity *E24 Physical Man-Made Thing*) was produced by the event of binding (CRM entity *E12 Production*) in the 14th century (CRM entity *E52 Time-span*). Then a later event (*E79 Part addition*) took place, where new endleaves (*E24 Physical Man-Made Thing*) were added in the 15th century (*E52 Time-span*). A similar activity (*E79 Part addition* of *E24 Physical Man-Made Thing*) occurred again in the 17th century (*E52 Time-span*). The CRM is inclusive enough to be able to model many, if not all, of the types of activities in conservation. A few comments on the CRM follow.

- Language: Researchers who get introduced to the CRM often feel that the names of the classes and the language of the CRM in general are alienating. For example, the term *Physical Man-Made Thing* sounds strange. Why not just use the term *object*? Following this initial stage, researchers realise that a different choice of words would exclude some possibilities. Not all human creations are objects - think about rock art in caves. Calling this entity *object* would exclude all rock art and similar items. So an important point to make is that the CRM language is strange, but it has

[9]In reality there are many more concepts and relationships, but this will suffice for this example.

been scrutinised thoroughly.

- Inheritance: CRM classes follow a hierarchical structure. So *Physical Man-Made Thing* is a subclass of *Man-Made Thing* which is a subclass of *Thing* which is a subclass of *Persistent Item* and so on. This is useful because everything that is true for a parent class is also true for a subclass. Therefore rules defined for parents apply to child classes automatically and save the repetitive task of redefining rules for each class.

- Event: Although not explicitly presented as such, it is reasonable to say that the CRM gives emphasis to the concept of the *Event*. *Event* is a nodal concept because it links the date of an activity, the location, the participating people (or organisations) and the object in use during the activity. By documenting an event the CRM allows the documentation of many associated entities which are important to describe the event and to match significant concepts in conservation documentation.

The CRM is the de-facto ontology for cultural heritage. It is a mature ontology with relatively minor revisions over the past years and it has also become an ISO standard [ISO, 2006]. The large variety of projects using the CRM is further proof of its maturity and importance. The CRM is a great asset in cultural heritage and an obvious choice for conservation, given the emphasis on activities and events. The CRM is a generic ontology and it does not include classes for specific domains. For example, the CRM includes the generic concept of *Condition State* but it does not include the possible values for it, such as *torn* or *faded*. Such expert classes are typically provided by a thesaurus and a good thesaurus to offer these classes is the *Getty Arts and Architecture Thesaurus* (AAT) as explained next.

Arts and Architecture Thesaurus [subheading]

The AAT is a project with a long history which has good coverage of expert classes in a range of cultural heritage domains. AAT has recently (2013) added many conservation terms in collaboration with the *Getty Conservation Institute* and it includes the classification terms used for the AATA Online abstracts. Therefore it is a reasonable choice for conservation documentation in combination with the CRM.

In the next section (Resource Description Framework [subheading]) I give a brief description of a well-tested technology (RDF[10]) which can act as a carrier for data expressed using CRM classes and AAT concepts (or in fact any other ontology and thesaurus). For RDF to work, the associated thesauri must conform to certain requirements. One of these is that each concept of the thesaurus has a unique web address, also called a *Uniform Resource Identifier* (URI)[11].

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

[11] For a detailed discussion of URIs see: http://www.w3.org/Addressing/URL/uri-spec.html

A recent announcement by the Getty [12] defines a timeframe (January 2014) by which the Getty vocabularies will be available through unique web addresses, thus making them technically suitable for use with RDF. For example, a reference to a concept of the thesaurus (e.g. *cinnabar*) will be made using its unique web address:

```
http://vocab.getty.edu/aat/300311452
```

The Getty vocabularies also include *The Getty Thesaurus of Geographic Names* (TGN), *The Union List of Artist Names* (ULAN) and *The Cultural Objects Name Authority* (CONA). The respective web addresses for these will be:

```
http://vocab.getty.edu/tgn/
http://vocab.getty.edu/ulan/
http://vocab.getty.edu/cona/
```

In the section about IIC website: a sample implementation [heading] I outline how these can be used in addition to the AAT to implement the IIC website. Before that, let us introduce RDF and explain how it can work with the CRM and a thesaurus.

Resource Description Framework [subheading]

In a previous section (Free text [subheading]) I emphasised the value of producing structured documents for conservation documentation. The main argument being that they can offer better search results than free text descriptions. However, in the section about schemas (Structured documents: schemas [subheading]), I explained why structured documents which are based on a schema, limit the capacity of search to only collections which conform to that schema. This is because different schemas hold different data and attempting to present different data comparatively is not applicable since comparisons are meaningful only among similar data. Therefore an attempt to unify all schemas under a single search framework should be based on what these schemas have in common.

RDF is based on the idea of a single common structure and it defines a simple framework to express data: a *triple*. It is widely accepted that triples can be used to express any piece of information and as such the structure of a triple is useful as a common denominator across various schemas. The data held in a document following a specific schema can be expressed in the form of triples and therefore separate collections can be queried together. Examples of data expressed as triples can be found on the British Museum website. The record of an Athenian acroterion with museum number 1843,0531.26 for website visitors can be found here:

```
http://www.britishmuseum.org/research/collection_online/
    collection_object_details.aspx?objectId=461640&partId
    =1
```

[12] http://www.getty.edu/research/tools/vocabularies/lod/index.html (last retrieved on 03-10-2013).

while the same record expressed in triples can be found here:

```
http://collection.britishmuseum.org/description/object/
    GAA6865.html
```

For example, the triple:

```
Subject: http://collection.britishmuseum.org/id/object/
    GAA6865
Predicate: http://collection.britishmuseum.org/id/crm/
    P52F.has_current_owner
Object: http://collection.britishmuseum.org/id/the-
    british-museum
```

indicates that the *current owner* of the object with database id GAA6865 (which internally corresponds to museum number 1843,0531.26) is the *British Museum*. The triple:

```
Subject: http://collection.britishmuseum.org/id/object/
    GAA6865
Predicate: http://collection.britishmuseum.org/id/crm/
    P57F.has_number_of_parts
Object: 1
```

indicates that the object consists of only one part (as opposed to, for example, a pyxis which may have a separate lid and therefore consist of two parts).

These triples are simply formed by putting three web references next to each other in the sequence of *subject*, *predicate* (or *property*), *object* (or *value*). Subject defines the concept (the object in our example) about which the statement is made. Property defines the characteristic of the concept. Object defines the value of the property and it can be another web reference or plain data (a number in our example). Each of these web references can refer to entities or concepts from ontologies and thesauri. In the above example the predicates are borrowed from the CRM, while the concepts are provided by the British Museum.

Although the British Museum implementation is exemplar, it is worth noting that the conservation records are not included in the list of triples, indicating the fact that free-text conservation records are difficult to process and that structured conservation records are needed. Recognising this gap and using the above resources and technologies, the next section introduces a system for publishing conservation data on the IIC website based on the CRM entities and with concepts from the AAT.

IIC website: a sample implementation [heading]

The content on the IIC website is divided into two main categories. The first one includes news items and announcements of events (including conferences) and the second one includes publications from various IIC journals (Studies in

Conservation, Reviews in Conservation, News in Conservation, Congress proceedings). This content is primarily in the form of free text and therefore it appears that there is little scope in publishing it as structured RDF triples. Indeed, much of the detail captured in free text would require huge effort to convert into structured documents. However, there is scope in isolating the more significant information across all content to produce a common search framework.

After consulting various stakeholders, reviewing the CRM and taking into account that the *event* is the core element in conservation documentation, it was decided that the information to be recorded for a piece of content will be[13]:

- *actor*, this matches the *E39 Actor* CRM class which includes individuals or groups with capacity to undertake action during an event. A conservator undertaking conservation work on an object would be an *actor*, as would be an institution organising a conference. This information allows the development of a list of participants in the various events published on the IIC website so visitors can search by participant/*actor*.

- *period*, this matches the *E52 Time-Span* CRM class which shows the period of time during which the event takes place. This allows the chronological arrangement of events.

- *place*, this matches the *E53 Place* CRM class which includes the location that the event took place. This information allows the development of a list of places where the various events published on the IIC website take place. The Getty TGN could be used as a reference for these locations.

- *thing*, this matches the *E70 Thing* CRM class which includes recognisable items of "relative stability". Museum objects and monuments can be considered as things and therefore any item present at the event (e.g. the object being conserved) would be included here. The Getty CONA could be used as a reference for these objects.

- *domain*, this is not intended to match a CRM class. The *domain* is used as a familiar way of indexing content on the IIC website for compliance with other conservation websites (see Domain index [subheading]). The AAT is used to provide the *domain* reference terms.

For the initial implementation of the website, any information entered in these fields is kept in an index with unique web addresses which can be referenced internally. Following the publication of the Getty vocabularies in an RDF-friendly format (as mentioned in Arts and Architecture Thesaurus [subheading]), this referencing will be done directly using the Getty addresses.

The two main categories of content (events and publications) include these fields and can be linked to them using CRM properties. For example:

[13]Readers with experience in the CRM may object with the naming of these pieces of content, as in some cases it clashes with the naming of the classes in the CRM. The choice of names was informed by the familiarity of the website editors with previous naming conventions.

- an *event* (*E5 Event*) is linked with *actor* (*E39 Actor*) with the property *P11 had participant*,
- an *event* is linked with a *thing* with the property *P12 occurred in the presence of,*
- a *publication* (*E89 Propositional Object*) is linked to a *thing* with the property *P67 refers to.*

Therefore individual pieces of content on the IIC website can be expressed as RDF using classes from the CRM and concepts from AAT and other Getty vocabularies. Moreover the fields discussed above can also be used as indexes for browsing and searching content on the website. *Domain* is discussed next.

Domain index [subheading]

Following the survey of conservation websites, part of which was described in Controlled vocabularies and thesauri [subheading], and after the analysis of the typology of content published on conservation websites (see the same section), an index of classification terms for IIC content was produced based on concepts included in the AAT. A selection of these can be found in the Appendix [heading]. This index was successfully tested for coverage with a large sample of existing content from the IIC website.

Software [subheading]

The IIC website has been built with open source software which means that replicating the implementation can be done without any licensing costs. In addition to the above requirements for publishing conservation data as triples, IIC relies on its website for a number of other services including receiving payments, managing the membership and sending out mass-mailings. In this article I am only focussing on the part of the implementation for publishing content, but the requirement for the other services was taken into account when choosing the website software.

The IIC website is built using the Drupal content management system[14]. Drupal was chosen for these main reasons:

- it is a modular system which allows new functionality to be added by installing new modules,
- it has built-in functionality for expressing content in RDF (core RDF module[15]) and a wealth of externally contributed RDF modules (e.g. RDFx[16]) which can be installed and make the core module more usable,
- it requires little training for website users to start publishing content.

[14] https://drupal.org
[15] https://drupal.org/node/1089804
[16] https://drupal.org/project/rdfx

Drupal makes no assumptions on the type of content which can be published on a website and allows flexibility regarding the type of fields linked to content. The model that Drupal is using to publish content as RDF triples is proposed inCorlosquet et al. [2009]. A Drupal *type of content* is considered as the RDF *subject*. A content type *field* can be considered as the RDF *predicate* and the value of the field can be considered as the RDF *object*. In practice, RDF data is delivered within web-pages as *RDFa* (a method of publishing RDF data inside ordinary HTML)[17]. An example of RDFa data delivered through a web-page is shown here[18]:

```
<div typeof="crm:E5.Event" about="/node/3147">
 <div class="field">
 <div class="field-label">Place:</div>
  <div class="field-item" datatype="xsd:string" property
      ="crm:P7.took_place_at">Vienna</div>
  <div class="field-item" datatype="xsd:string" property
      ="crm:P7.took_place_at">Austria</div>
 </div>
</div>
```

When this code is rendered by the browser (processed and displayed on screen) it will appear as:

```
Place: Vienna Austria
```

Inside the HTML code as shown above is the RDF data. This implementation of conservation data with Drupal is recommended as it allows publishing RDF data without having to programme (often elaborate) templates. The contributed module *RDFx* offers a usable interface where Drupal data can be mapped to entities of an ontology. This concentrates the focus of the website development to ontological questions and relieves the burden of solving technical problems (such as coding syntax). This implementation worked well for the IIC website, but it may be less suitable for detailed conservation content with more complex semantic relationships. The flexibility of Drupal means that the system could deliver such complexity but the current version of the RDFx module would not be adequate. If more complex RDF data is to be published then programming templates on a per-content type case would be required.

Conclusions and Future work [heading]

In most cases conservation documentation gives emphasis on describing objects as static items. In a previous section () I explained how this is limiting the scope of documentation systems. Recent work in conservation documentation

[17] More information about RDFa can be found here: http://www.w3.org/TR/xhtml-rdfa-primer/

[18] For the requirements of this illustration the code has been simplified by removing the internal referencing to the geographical locations of "Vienna" and "Austria". This referencing will be revised once the AAT vocabularies are available in an RDF-friendly formal.

indicates that an event-centric approach may be more appropriate for capturing the history of objects including their past treatments. The same principle can be applied to the organisation of online conservation content and a case study from the IIC website has been successful in adopting this principle. Wider adoption would lead to more case studies being investigated to fully appreciate the benefits of one approach over the other.

Semantic Web technologies are rarely used in conservation documentation. There are few examples of attempts to enrich conservation records semantically as explained in Previous attempts [heading], but in most instances conservation websites rarely offer any semantically-rich content. Instead they rely on familiar language for organising content which includes compound terms, mixing different aspects of records and therefore making machine searching difficult. Recommendations from the field of knowledge organisation suggest the use of controlled vocabularies and thesauri for organising content. A thesaurus with good coverage in the field of conservation is the AAT. A survey of popular conservation websites showed that in terms of the underlying concepts many of the categories used are identical and that it would be possible to propose a common set of concepts to organise content. This set of concepts can be extracted from the AAT and is presented in the Appendix [heading]. Organising the content of conservation websites based on AAT concepts would assist machine searching in conservation in the long term.

The CIDOC-CRM is the defacto standard for expressing relationships between concepts in cultural heritage and conservation and although museums and other cultural organisations have adopted it for several documentation systems, it is rarely used in conservation documentation. The case-study of the IIC website showed that there are currently mature tools which can be used to deliver conservation content online using CIDOC-CRM. Drupal is a good solution because it does not necessarily require technical knowledge of semantic technologies, and allows the user to focus on the underlying concepts in relation to their conservation expertise.

The RDF implementation in Drupal which can be used to express semantic content is a fixed model (Drupal content type $\rightarrow$ RDF subject, Drupal field $\rightarrow$ predicate, Drupal field data $\rightarrow$ object) and may dictate the structure of content to match CIDOC-CRM relationships. There may be benefit in considering alternative models where this correspondence is not fixed especially for websites which wish to publish more detailed data.

As explained in section there is value in querying conservation records across collections and in order to do that online, the adoption of semantic web technologies is essential. It is appropriate to widen the discussion on conservation documentation now so that such querying will soon become possible as new conservation records are published online in suitable formats.

Appendix [heading]

`Activities Facet`

--Functions (Facet)

----functions (activities)

------<functions by general context>

--------<analytical functions>

----------authentication

----------research (function)

--------<information handling functions>

----------collections management

----------documentation (activity)

------------conservation documentation

--------<organizational functions>

----------management

------------disaster planning

--------maintenance

----------preventing

------------pest control

----------environmental control

--Disciplines (facet)

----disciplines

------social sciences

--------<history and related disciplines>

----------history (discipline)

------------conservation history

--------education

----------<education by subject>

------------conservation education

--------law (discipline)

----------conservation law

------<science and related disciplines>

--------science

----------conservation science

----------natural sciences

------------physical sciences

--------------physics

----------------electronics

------------------electronic engineering

------<cross- and interdisciplinary fields>

--------conservation

----------<conservation by collection type>

------------archive conservation

------------archaeological conservation

------------architectural conservation

------------natural history conservation

------------industrial heritage conservation

----------<conservation by activity>

------------preventive conservation

------------object conservation

------humanities

--------<arts and related disciplines>

----------<arts-related disciplines>

------------art history

--------philosophy

----------ethics (philosophy)

------------conservation ethics

--Processes and Techniques

----<processes and techniques>

------<processes and techniques by specific type>

--------<object-making processes and techniques>

----------basketmaking

--------<additive and joining processes and techniques>

----------<surface covering processes and techniques>

------------metallizing

--------------gilding

Associated Concepts Facet

--Associated Concepts

----<scientific concepts>

------<physical sciences concepts>

--------<earth sciences concepts>

----------<weather and related phenomena>

------------climate

----<social science concepts>

------<economic concepts>

--------<industry (economic concept)>

----------tourism

----<technology and related concepts>

------technology

--------information technology

----------<computer networking concepts>

------------born digital

----<functional concepts>

------storage

----<transportation and related concepts>

------transportation

Physical Attributes Facet

--Color (Facet)

----<color and color-related phenomena>

------color (perceived attribute)

--------<color-related attributes>

----------multicolored

------------polychrome

--Attributes and Properties

----<attributes and properties>

------<attributes and properties by specific type>

--------physical properties

----------waterlogged

Materials Facet

--Materials (Hierarchy Name)

----<materials by composition>

------inorganic material

--------rock

----------<rock by form>

------------stone (rock)

--------glass (material)

--------metal

--------clay

----------<clay products>

------------<ceramic and ceramic products>

--------------ceramic (material)

----------------porcelain

------organic material

--------plastic (organic material)

--------resin (organic material)

----------fossil resin

------------amber (fossil resin)

------<combination inorganic/organic material>

--------<combination inorganic/organic animal material>

----------bone (material)

------------<bone by form>

--------------<tooth and tooth components>

----------------<tooth components>

------------------dentin

--------------------ivory (tooth component)

----<materials by function>

------coating (material)

--------<coating by form>

----------paint

--------<coating by composition or origin>

----------lacquer (coating)

----------enamel (fused coating)

------colorant (material)

--------dye

------photographic materials

--------photographic film (photographic materials)

----<materials by form>

------<materials by physical form>

--------<fiber and fiber products>

----------<fiber products>

------------paper (fiber product)

----<materials by origin>

------<biological material>

--------animal material

----------<collagenous material>

------------skin (collagenous material)

----------<processed animal material>

------------leather

--------plant material

----------<wood and wood products>

------------lichens

------------wood (plant material)

Objects Facet

--Built Environment (Hierarchy Name)

----Single Built Works (Hierarchy Name)

------<single built works (Built Environment)>

--------<single built works by general type>

----------structures (single built works)

------------buildings (structures)

--Furnishings and Equipment (Hierarchy Name)

----Furnishings (Hierarchy Name)

------furnishings (artifacts)

--------<furnishings by form or function>

----------Costume (Hierarchy Name)

------------costume (mode of fashion)

----------frames (furnishings)

----------furniture

--Object Genres (Hierarchy Name)

----<object genres (Guide Term)>

------<object genres by material>

--------textiles

------<object genres by location, context or origin>

--------ethnographic objects

--------archaeological objects

--Visual and Verbal Communication (Hierarchy Name)

----Visual Works (Hierarchy Name)

------<visual works (Guide Term)>

--------<visual works by medium or technique>

----------electronic images

------------digital images

----------paintings (visual works)

------------<paintings by location or context>

--------------mural paintings (visual works)

--------------stained glass (visual works)

----------photographs

----------sculpture (visual work)

----Information Forms (Hierarchy Name)

------<information forms (Guide Term)>

--------<information artifacts>

----------<information artifacts by function>

------------<identifying artifacts>

--------------labels (identifying artifacts)

--------------seals (artifacts)

----------<information artifacts by physical form>

------------books

--------<document genres>

----------<document genres by function>

------------<identifying markings and symbols>

--------------marks (symbols)

```
----------<document genres by form>

------------<graphic document genres>

--------------cartographic materials

----------------maps
```

References

M. Baca. Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & classification quarterly*, 36(3-4):47–55, 2003.

C.-H. Chang, M. Kayed, R. Girgis, and K.F. Shaalan. A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411 –1428, October 2006. ISSN 1041-4347. doi: 10.1109/TKDE.2006.152.

Stephane Corlosquet, Renaud Delbru, Tim Clark, Axel Polleres, and Stefan Decker. Produce and consume linked data with drupal! In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009*, volume 5823, pages 763–778. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-04929-3, 978-3-642-04930-9. URL `http://data.semanticweb.org/conference/iswc/2009/paper/inuse/101/html`.

Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, editors. *Definition of the CIDOC Conceptual Reference Model*. December 2011.

Alex Dawson and Susanna Hillhouse, editors. *Spectrum 4*. Collections Trust, May 2011.

ISO. *Information and documentation : a reference ontology for the interchange of cultural heritage information*. Number ISO 21127 in International standard. ISO, Geneva, 2006.

Arno Mittelbach. Roma: generating validators for the TEI. URL `http://www.tei-c.org/Roma/`.

Nicholas Pickwoad. The condition survey of the manuscripts in the monastery of saint catherine on mount sinai. *The Paper Conservator*, 28(1):33–61, 2004. ISSN 0309-4227. doi: 10.1080/03094227.2004.9638640. URL `http://www.tandfonline.com/doi/abs/10.1080/03094227.2004.9638640`.

Heather Ravenberg. *A data model to describe book conservation treatment activity*. MPhil, University of the Arts London, London, 2012.

Jenn Riley. Seeing standards, 2009. URL `http://www.dlib.indiana.edu/~jenlrile/metadatamap/`.

Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1):233–272, 1999. ISSN 0885-6125. doi: 10.1023/A:1007562322031. URL `http://www.springerlink.com/content/m23n8197vg924t51/abstract/`.

TEI Consortium, editor. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* 5 edition, October 2007.

A. Velios and N. Pickwoad. An optimised workflow for large-scale condition surveys of book collections. In M. Driscoll and R. Mosesdottir, editors, *Care and Conservation of Manuscripts*, volume 11, pages 269–290, Copenhagen, 2009. Museum Tusculanum Press. ISBN 9788763530996.

Jiemin Zhang, April Webster, Michael Lawrence, Madhav Nepal, Rachel Pottinger, Sheryl Staub-French, and Melanie Tory. Improving the usability of standard schemas. *Information Systems*, 36(2):209–221, April 2011. ISSN 0306-4379. doi: 10.1016/j.is.2010.08.005. URL `http://www.sciencedirect.com/science/article/pii/S0306437910000827`.

Acknowledgements

The author thanks Martin Doerr and Maria Theodoridou from the Foundation of Research and Technology Hellas (FORTH) for their pioneering work and expert advice. The redevelopment of the IIC website was made possible with generous funding from the Getty Foundation and is supported by Ligatus and the University of the Arts London.

Images

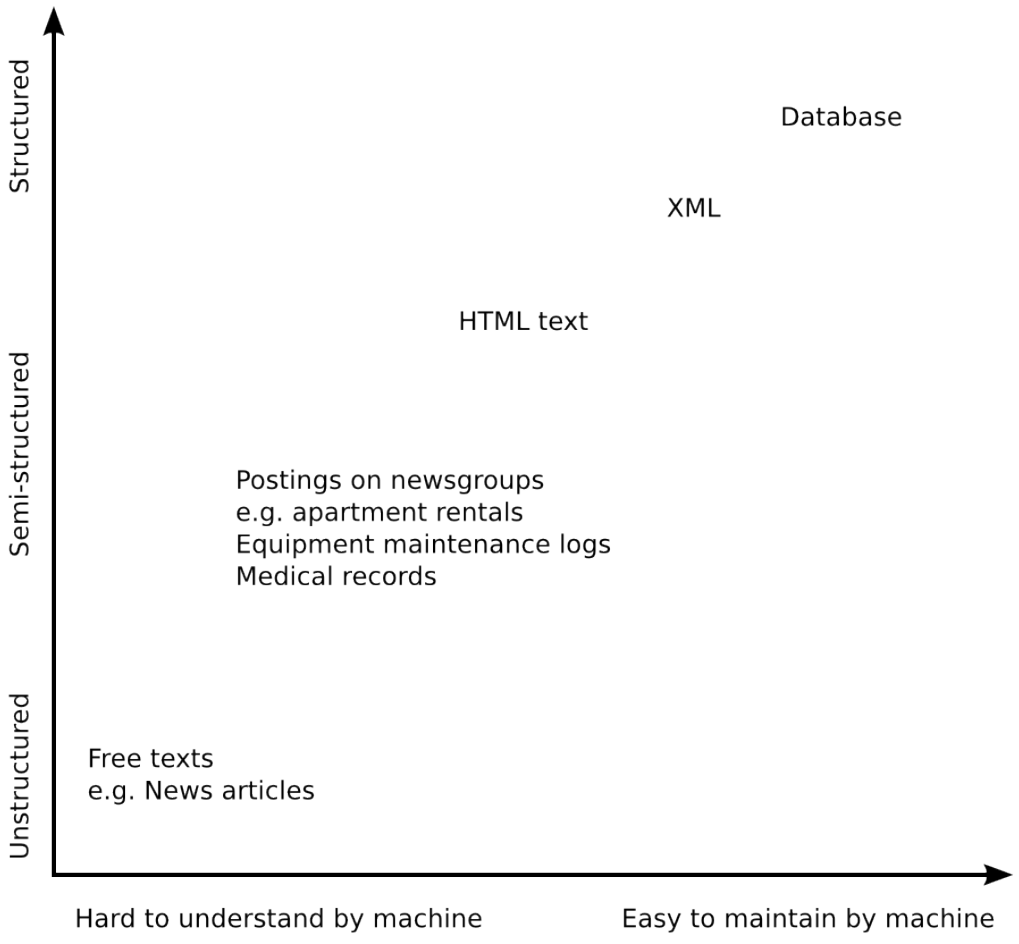

Figure 1: Structured text is easier to process than unstructured (after [Chang et al., 2006, § 3.1]).

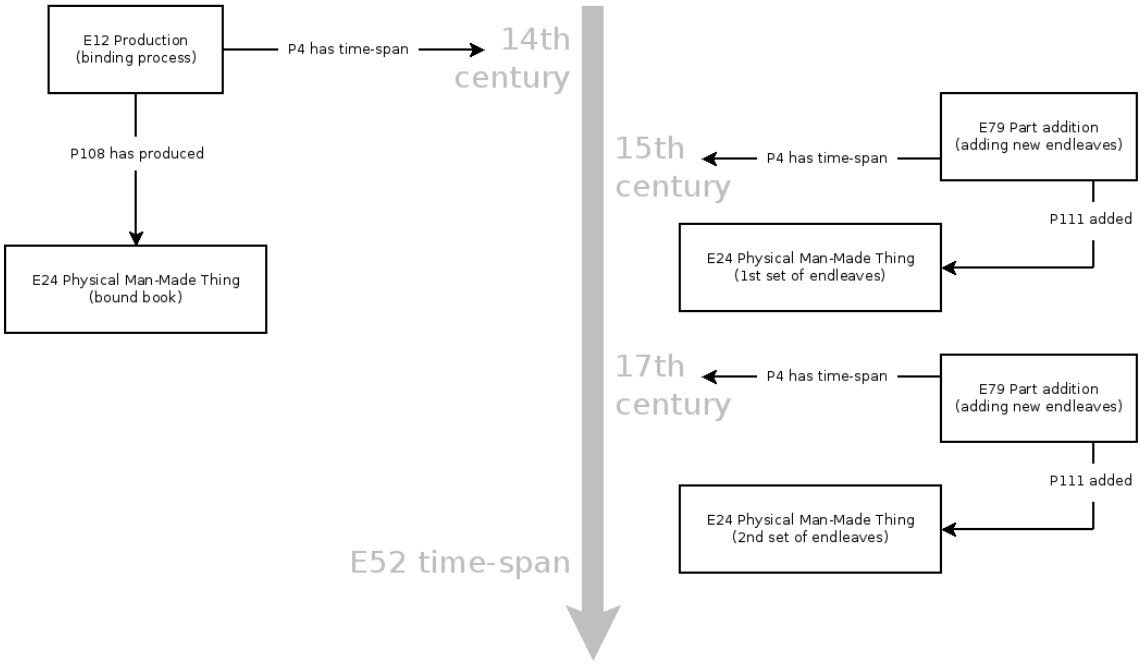

Figure 2: Using CRM concepts to map multiple additions of endleaves to a binding.