Archeologia e Calcolatori
20, 2009, 111-128

REPRESENTING KNOWLEDGE IN ARCHAEOLOGY: FROM CATALOGUING CARDS TO SEMANTIC WEB

1. Introduction

Representing knowledge is the basis of any catalogue. The Italian Catalogue was based on very valuable ideas, developed in the late 1880s, with the basic concept of putting objects in their context. Initially cataloguing was done manually, with typewritten cards. The advent of computers led to some early experimentations, and subsequently to the definition of a more formalized representation schema. The web produced a cultural revolution, and made the need for technological and semantic interoperability more evident. The Semantic Web scenario promises to allow a sharing of knowledge, which will make the knowledge that remained unexpressed in the traditional environments available to any user, and allow objects to be placed in their cultural context.

In this paper we will briefly recall the principles of cataloguing and lessons learned by early computer experiences. Subsequently we will describe the object model for archaeological items, discussing its strong and weak points. In section 5 we present the web scenario and approaches to represent knowledge.

2. Cataloguing: history and principles

The Italian Catalogue of cultural heritage has its roots in the experiences and concepts, developed in the late 1880s and early 1900s, by the famous art historian Adolfo Venturi, who was probably one of the first scholars to think explicitly in terms of having a frame of reference to describe works of art, emphasizing as the main issue the context in which the work had been produced. In 1964, the famous art historian Giulio Carlo Argan promoted a study group (under the Ministry of Education and CNR - National Research Council) and proposed the establishment of a specific body in charge of cataloguing. The Italian Cataloguing Institute (ICCD) was officially born in 1975, under the direction of prof. Oreste Ferrari.

The key principle guiding the cataloguing process is *knowledge*, about the specific object but also about all the other issues that can help in understanding the complex semantic relationships the object has with issues pertaining to other disciplines. The full knowledge of the historical, political, religious context is in many cases the only way to fully understand the value and the

message of an object. However, such knowledge is almost invariably the property of the scholars and experts, and rarely is made explicit to the others. As a result, many of us can only perceive a minimal part of what is the real value and meaning of works of art. Anyone who has had the experience of visiting a museum accompanied by an expert who can describe details about the cultural context in which an object was made, can easily understand the difference between this kind of visit and a conventional one.

3. Early experiences and cataloguing cards

Computer applications in the area of management of cultural heritage date back to the 1960s. In many cases, the approach taken to storing catalogue information was very similar to the one used by librarians. The basic idea was to describe objects with "cataloguing cards" where information was organized in several semantically consistent sections, describing, for example author, period, excavation data, subject, historical and critical notes. The first organisation of the Italian Catalogue was based on a manual approach, where each object was described by a typewritten card. The basic ideas were very valid and all subsequent work has been greatly influenced by the intellectual efforts that led to the definition of the fundamental principles of the cataloguing rules. The most important issues were:

– identification of a reduced set of different cards, corresponding to different types of objects (art objects, archaeological objects, drawings, architecture, gardens, historical centres, etc.);
– grouping of the information in several very general categories, like author, location, material, historical info, etc.;
– topological arrangement of the catalogue cards.

On the other hand, it should be pointed out that the cards were conceived for human usage, and therefore the various fields were to be filled in by scholars, on the basis of their specific competence in the particular subject, following some general rules. As a matter of fact, it was expected that the cards would be used by another scholar, who would be able to understand the semantics of the content of the fields, and identify any inconsistencies, or interpret them correctly.

This approach would, in principle, fit all the needs, but there were some important points missing. First of all, cultural heritage is far more complex than a library, as it is intrinsically highly interdisciplinary. Secondly, cards are compiled by humans and intended for humans, who can read, understand, infer, connect, and reason about their content. In this light, the fact that information is not highly formalized is not a problem: free text is widely used and concepts can be easily expressed. Third, the guiding principle was "one

card for each object", regardless of its complexity, even if some objects can be seen and described as single items, while others have more complex structures and relationships with other objects.

The library approach of free text cards, with some more formalized items, was proven to be somehow semantically ambiguous with free text information potentially inconsistent. This fact became evident when, in the early 1970s, some more broad range experiments were conducted on the use of computers to store catalogue data. The first and quite natural approach was just to transform the paper cataloguing card into an electronic document, relying on the features and effectiveness of the Information Retrieval Systems (IRS). The assumption was that it would be easy, for scholars as well for casual users, to enter some words the system could find in documents and return appropriate records. In spite of the initial enthusiasm and the claims of the vendors, some of the initial results were both unexpected and very disappointing; as the quantity of data increased, it became evident that both precision and recall were not achieving the expected optimum[1]. In fact, when indexing documents, IRSs use a list of non significant terms (stopwords) to avoid indexing of useless or non selective words. In some systems the stopwords were system-wide, while in others they could be linked to specific parts of the document. In any case, the problem of homographs produced disappointing results. Just as an example, in Italian the capital "i" ("I") is used both as article as well as the Roman number denoting the ordinal number one, as in the expression "I secolo" to denote the first century. Depending on the characteristics of the system and the list of stopwords, a search for objects dated to the first century returned either zero documents or almost all objects. Another disappointing effect was caused by the adoption of a *flat* model, where a single document was describing the object, irrespectively of its complexity. This made it impossible, for example, to find objects made of several different materials, or having components with a different style or date.

It is worthwhile to note that while the first effect was related to the software features, the second one depended on the rule of having "a card for each object". In general, lack of precision was caused by the poor, flat structure of the cataloguing card, while the absence of a controlled language was the origin of poor recall.

4. The object model and the RA card

The automation of the Italian Catalogue started in the mid 1970s, and during the first experiences only the IRS approach was used, mapping

[1] It should be noted that the effectiveness of IRS is basically measured in terms of Precision and Recall, where Precision is the ratio between the number of returned significant documents and the total number of returned documents, and Recall is the ratio between the number of returned significant documents and the number of significant documents in the whole document collection.

paper cards to electronic documents. In addition, many experiments were conducted in order to determine how to map the different cataloguing cards (for example, the art objects and the archaeological artefacts) to different electronic document structures. Even if these attempts paid more attention to the structuring of documents, they were still tied to the software selected as IRS, and results were mostly disappointing. It soon became evident that the unsatisfactory results could not be ascribed to the particular systems (every product exhibits some strong and some weak points), but that a rethinking of the entire cataloguing schema was necessary, keeping in mind the constraints imposed by the automated treatment of the information. Around 1984 it was decided to take a different approach[2], which had the following guidelines:

– higher degree of formalization;
– independence from software and hardware platforms;
– unified model for the different kinds of objects, and therefore, as far as possible, a "unique schema".

The flat model which was the basis of the cataloguing card was reconsidered. The main issue was therefore the definition of a model for the objects, trying to abstract from the differences which distinguish the different objects, as they are seen according to the various disciplines to which they are related. For example, a vase or a brooch are seen and described differently depending on whether they are an archaeological artefact or an art object. Another relevant issue considered was the distinction among different types of objects, taking into account their intrinsic properties. The definition of the data structure started from the identification of the central role played by the object. Initially, a wide variety of object types was identified: single object, series, fragment, part of, etc. After a while, it was realized that this kind of specialization was too complex, and would in fact constitute a barrier between the cataloguer and the user. Finally, it was agreed to define a classification schema based on three different types of objects:

– *simple object*: is an object such that all its attributes are pertinent to the whole object, and cannot be separated into smaller components, exhibiting different relevant properties (different materials, different epoch, different authors, etc.) which may themselves be considered cataloguing objects;
– *complex object*: may be either a simple object whose parts are physically or conceptually separable and exhibit some interesting peculiarities as cataloguing objects, or a set of objects which may be referred to by a specific name;

[2] These were the years when database conceptual modelling was mature, the huge and heavy hierarchical DBMSs were being replaced by the more theoretically sound relational databases, and the object oriented approach was getting attention from researchers. The web, where technical interoperability is a familiar feature, was not yet invented.

– *aggregation of objects*: occurs when a set of objects can be considered as a whole on the basis of some conceptual perspective, and this set has properties which are useful to express, but no name exists which identifies the aggregate.

Components of a complex object may be either simple or complex objects, and so may the aggregate objects. It is worthwhile noting that a specific object belongs to the different categories only on the basis of the quantity and the type of information: no list exists that specifies if a particular kind of object must be considered simple or complex or aggregate. The proposed model only established a classification model, that is, the type of relationships that must be specified between the objects (a component of a complex object is an object itself), and the criteria inherent to the properties. The model implicitly assumed the existence of relationships between objects (as will be discussed below).

The approach used was essentially based on the standard conceptual database design methodologies. As it is well known, the conceptual model is independent from software and hardware environments, and the most popular approach at the time was based on the Entity-Relationship model. The first step was the identification of the "basic" entities, like object, author, location, and so on. The identification of the relationships between these entities was taken as the second step. This process lead to a simple, consistent model, where the object was playing a central role. In addition to the "fundamental" entities and relationships, which are intrinsic to the representation of the real world, there were some "minor" relationships and entities, as those accounting for the name variants. The analysis of these last and similar problems pushed us to the definition of some "authority files" as the only means to normalize the vocabulary, and to keep the data consistent.

Even though the Entity-Relationship model proved very effective in modelling and representing knowledge, cultural traditions and mental attitude forced them to switch to the conventional "cataloguing card" format. From this point of view, the following choices were made:

– the information has been subdivided into small, semantically well defined, chunks;
– these chunks may be either a field, or a subfield of a structured field;
– each field may be defined as simple or structured;
– each field may be defined as repeating or non-repeating;
– each subfield may be a repeating or non repeating subfield;
– fields, either structured or unstructured, may be grouped into "paragraphs" in order to allow multiple occurrences of a set of fields[3]. An example of a cataloguing card follows:

[3] As a curiosity, we may recall that fields are identified by three letter codes, subfields by four letters codes, and paragraphs by two letters codes. The length of codes was a legacy of the IRS used in the first experiences, which allowed for field labels of a maximum length of four letters.

```
CD:
TSK: RA
[…]
OG:
OGT:
OGTD: parete affrescata
LC:
PVC:
PVCP: NA
PVCC: ERCOLANO
LDC:
LDCT: casa
LDCN: V 15 (DEL BICENTENARIO)
LDCS: 13 (tablino); parete N
[…]
DA:
DES:
DESS: Zona mediana rossa con tre pannelli riquadrati da bordi
di tappeto: centrale (giallo in orig.) con quadro (Pasifae e
Dedalo), laterali con medaglioni (Bacco a s.; baccante a d.)
separati da fasce nere con grottesche, cornice a ovoli. Fregio
nero: tre pannelli con scene di amorini in caccia con scudo e
sfondo naturalistico a s., con cervo e cane in lotta al cen-
tro, con amorino su cavallo alato a d., separati da riquadri
con maschere tragiche. Zona superiore rossa con architetture
(quasi illeggibile).
[…]
```

At first glance, the proposed model may appear just another "flat file" schema, with a large number of fields, but anyone familiar with database design methodologies will easily recognize that, generally speaking, entities have been mapped on to paragraphs, (multivalued) attributes on to (repeating) fields, aggregate attributes on to structured fields. It is also evident that the identification of a sequence of fields, with the characteristics of being repeatable and/or groupable, and references to "authority files", may be seen as the "linearization" of a non linear text. Last, but not least, an effort was made to maintain consistency between different cultural areas, so that semantically equivalent fields are identified by the same tag.

The proposed model was a trade-off between the very specific requirements posed by the academic and research communities, asking for exhaustive information, and requirements posed by the administrative needs of having a model which could be adapted to a large variety of objects, despite their differences. The object model could also be seen as a way to *represent knowledge*, considering that, at the time it was defined, thesauri, authority files and dictionaries were not available for all the fields, and we were forced to use structured fields to represent knowledge that could be otherwise coded

in an appropriate faceted thesaurus. The object model could also be seen as an *interoperable model to collect information*.

The object model put a great emphasis on the definition of the different types of *relationships* between objects, leading to the definition of the different types of objects: simple, complex, aggregate. It has to be noted that a component of a complex object is seen as an object itself, and inherits some properties from the "father object". The aggregation of objects leads to the compilation of a cataloguing card that accounts for the general properties of the aggregate. In conclusion, we may envisage two different types of relationships between objects: a "vertical" relationship (complex objects) and a "horizontal" relationship (aggregates). As a consequence of experiences conducted in the following years, it was agreed to add the possibility of explicitly coding some semantically richer relationships. This deep and articulated fragmentation of information had the positive effect that information can be more accurately controlled, and errors are less probable, while fragments can be recombined to return more aggregated info. The guidelines followed in the definition of the standards are reported in Papaldo *et al.* 1986 and in Signore 1986. A complete definition of these standards may be found in the publications of the ICCD: D'Amadio, Simeoni 1989; Massari *et al.* 1988; Papaldo *et al.* 1988; Parise Badoni, Ruggeri 1988.

4.1 *Strong and weak points*

The definition of the object model and the corresponding cataloguing card exhibited several strong points. First of all, it made it possible to overcome the difficulties arising from the adoption of an approach tied to the technology to be used for storage and retrieval, while it was usual, in the mid 1980s, to implement models based on the software selected for the application management, with additional costs to face when the operating environment changed. Instead, everyone had total freedom in implementing data entry applications and sharing or exchanging data on the basis of the well defined model, thus safeguarding investments[4] (Signore 1993, 1994). Secondly, it was a big step forward, since it put the Italian Catalogue at least at the same level as other

[4] In 1986, the Italian government funded a Lit. 600.000.000.000 (approximately € 310 million) initiative, whose principal aim was the application of new technologies in the field of cultural heritage management. The initiative took the name of "giacimenti culturali", as it was assimilating the cultural heritage to other types of resources to exploit, like oil or coal. After a call for proposal, some 39 projects were approved and financed. Among them, 31 were concerned in some way with the cataloguing of works of art. No guidelines were imposed as far as the technological (hardware and software) environment was concerned, the only constraint was that the results of the projects should be made available to the central administration. The object model defined by the ICCD played a driving role as a standard at the *conceptual level*, and was included as a constraint in the contracts signed by the firms which were conducting the projects. The exchange format was easily defined in a couple of hours!

more advanced initiatives in other countries. Third, and perhaps most important, the model has been proven to be long lasting. The present schema in XML is not very different from the original one, and only a few adjustments have been made to the original structure, on the basis of experience.

However, we must recall some of the weak points, namely:

– the perception of a rigid schema with an excessive fragmentation of information;
– appropriate instruments to fill in the fields were missing;
– updating of controlled dictionaries, authority files and thesauri was slow and complex.

As a matter of fact, however, the number of fields was not so high, compared with the complexity of information to be represented, and the schema was thought of as an "extensible schema", as we were well aware that new needs could lead to its improvement. Lacking appropriate software to fill in the data was instead an obstacle for the immediate adoption of the model. Difficulties arising from the updating of controlled dictionaries and thesauri must be considered not a limit of the approach itself, but a consequence of the intrinsic complexity of organizing knowledge, emphasized by the lack of the cooperative tools we use today.

We must, however, mention some limitations of the approach itself. First, the approach remained centred on the traditional view of "one card for a single object", even if the object was modelled in a more elaborate way than usual. This "object centred vision" is the origin of redundant (and potentially inconsistent) information, like author or excavation data, and does not permit the representation of semantically complex associations. We must stress that these limitations were not in the original design, which had its roots in the database conceptual modelling, but is a consequence of the representation as a "cataloguing card", which puts too much emphasis on the object and imposes a linearization of the schema, that can exploit the (binary) associations between the object and other entities, but cannot express the existing interdisciplinary associations. As a consequence, the knowledge of the expert is not formally expressed and remains unavailable to the user. Even worse, linking with other disciplines was substantially impossible, and could be done only by expert users.

After the definition of the model, between 1985 and 1990 many experiments were conducted for the purpose of checking the correctness of the schema. At the same time, thesauri were created; of these, it's worth remembering the thesaurus of ecclesiastic furniture and the historical/geographical data bank, the pilot project from which the TGN was born (Papaldo, Signore 1989). Unfortunately, in 1990, when Oreste Ferrari retired, the activity was stopped, and the full project was never completed. As a consequence, the data entry and

exchange format was definitively taken as return format, and less effort was dedicated to the creation of thesauri and knowledge representation.

5. Towards Semantic Web

5.1 *The "web revolution"*

The web exploded in the mid 1990s, and was the origin of a true revolution of the traditional means of accessing information. Among the many characteristics of the web, we must recall a complete transformation of traditional methods of accessing information (Coyle 2007). In the past, users were accessing information starting from the "official repositories", like libraries, museum catalogues, and so on, while now they almost invariably start from a generic query on the web, and then follow the links, looking for the relevant information. As a consequence, the role of central repositories is much less important, as the web architecture is fully decentralized, and two issues emerge: the technical interoperability, which is granted by the web protocols, and the *semantic interoperability*, which could enable us to combine knowledge available from different sources. The latter is the most relevant issue, as it requires the representing, exporting and sharing of knowledge.

5.2 *Levels of knowledge representation*

The degree of formalization of concepts and their relations varies considerably among different domains of knowledge. At the lower end one finds lexicons and simple taxonomies, at the middle level one might place thesauri, at the high end of formalization of knowledge there are axiomatized logic theories. Such theories include rules to ensure the correct formulation and logical validity of statements expressed in the language of the scientific discipline (*Digicult* 2003).

According to Sheth and Ramakrishnan (2003) semi-formal ontologies[5], defined as ontologies that do not claim formal semantics and/or are populated with partial or incomplete knowledge, can be significantly smaller, especially for the ontology population effort, compared to that required for developing formal ontologies or ontologies with more expressive representations. Semi-formal ontologies have provided good examples of both value and

[5] The term ontology was taken from philosophy, where it denotes a specific subfield, namely the study of the nature of existence. It is the branch of metaphysics concerned with identifying, in the most general terms, the kinds of things that actually exist, and how to describe them. The observation that the world is made up of specific objects that can be grouped into abstract classes based on shared properties is a typical ontological commitment. More recently the term ontology has become relevant in the Knowledge Engineering community, acquiring a specific technical meaning, rather different from the original one. In fact, instead of "Ontology" we speak of "an ontology". Several different definitions of ontology exist, highlighting different aspects.

utility in meeting several challenges; especially that of information integration. One key reason is that of the need to accommodate partial (incomplete) and possibly inconsistent information, especially in the assertions of an ontology. Real world applications often can be developed with very little semantics or with compromises with completeness and consistency required by more formal representations and inferencing techniques ("a little semantics goes a long way").

Hierarchical classification systems and structured vocabularies do not lend themselves easily to rich inter-linking of conceptual "trees". A major step further in this direction is the "CIDOC object-oriented Conceptual Reference Model" (CRM). This provides an ontology of 81 classes and 132 unique properties, which describes in a formal language concepts and relations relevant to the documentation of cultural heritage[6]. CIDOC CRM is a formal ontology for cultural heritage information specifically intended to cover contextual information. It can be used to perform reasoning (e.g. spatial, temporal).

5.3 *The Dublin Core standard*

As clearly explained by Baker 2000, Dublin Core is often presented as a modern form of *catalogue card*, a *set* of elements (and now qualifiers) that describe resources in a complete *package*. Sometimes it is proposed as an *exchange format* for sharing records among multiple collections. A founding principle is that "every element is optional and repeatable". Strictly speaking, a Dublin Core element or qualifier is a unique identifier formed by a name (e.g., creator) prefixed by the URI of the namespace in which it is defined, as in http://dublincore.org/documents/dces/#creator/. In this context, a namespace is a vocabulary that has been formally published, usually on the web; it describes elements and qualifiers with natural-language labels, definitions, and other relevant documentation. The fifteen elements of the Dublin Core element set are the defining feature of Dublin Core as a language. In their short form, the elements are dc:title, dc:creator, dc:subject, dc:description, dc:publisher, dc:contributor, dc:date, dc:type, dc:format, dc:identifier, dc:source, dc:language, dc:relation, dc:coverage, and dc:rights. These correspond to fifteen broadly defined *properties* of resources that are generally useful for searching across repositories in multiple domains. Dublin Core is, in effect, a class of statements of the pattern *Resource has property X*, where "resource" is the implied subject; followed by an implied verb ("has"); followed by one of fifteen properties from the Dublin Core element set; followed by a property value, an appropriate literal such as a person's name, a date, some words, or

[6] The CIDOC CRM has been accepted as working draft by ISO/TC46/SC4/WG9 in September 2000. Since 9/12/2006 it is official standard ISO 21127:2006. See http://cidoc.ics.forth.gr/ for details.

a URI. For example: *Resource has dc:creator "Oreste Signore"*, and *Resource has dc:date "2009-04-01"*. Optional qualifiers may make the meaning of a property more definite.

5.4 *The Semantic Web*

To understand the Semantic Web framework it should be recalled that the web must be seen as a Universal Information Space, navigable, with a mapping from URI (Uniform Resource Identifier) to resources. For the Semantic Web to function, computers must have access to a structured collection of information and a set of inference rules that they can use to conduct automated reasoning. The challenge of the Semantic Web is therefore to provide a language that expresses both data and rules for reasoning about data and that allows rules from any existing knowledge-representation system to be exported onto the web.

The foundation of Semantic Web is the Resource Description Framework (RDF[7]) based upon a model for representing named properties and property values. The RDF model draws on well-established principles from various data representation communities. RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources and an RDF model can therefore resemble an entity-relationship diagram. The RDF data model is a syntax-neutral way of representing RDF expressions. The basic data model consists of three object types:

– *Resources*. All things being described by RDF expressions are called resources. Resources are always named by URIs plus optional anchor ids. Anything can have a URI; the extensibility of URIs allows the introduction of identifiers for any entity imaginable.
– *Properties*. A property is a specific aspect, characteristic, attribute, or relation used to describe a resource. Each property has a specific meaning, defines its permitted values, the types of resources it can describe, and its relationship with other properties. Each property is identified by a name, and takes some values.
– *Statements*. A specific resource together with a named property plus the value of that property for that resource is an RDF statement. These three individual parts of a statement are called, respectively, the *subject*, the *predicate*, and the *object*. The object of a statement (i.e., the property value) can be another resource or it can be a literal; i.e., a resource (specified by a URI) or a simple string or other primitive datatype defined by XML. A set of properties referring to the same resource is called description.

[7] See http://www.w3.org/TR/rdf-primer/ for an introduction and reference to other documents.

We can diagram an RDF statement pictorially using directed labelled graphs (also called "nodes and arcs diagrams"). In these diagrams, the nodes (drawn as ovals) represent resources and arcs represent named properties. Nodes that represent string literals will be drawn as rectangles. The power of RDF is that everything but the literals is identified by URI, and statements can predicate anything on anything, regardless of where they are located in the web. Therefore, the knowledge base is universal and worldwide. It is important to stress that the Semantic Web does not require that all the knowledge be migrated into RDF, it is sufficient that the existing knowledge, stored in databases, spreadsheets, documents, be mapped onto RDF graphs, so that it can be shared and queried by Semantic Web applications.

Semantic Web is a hot research topic, and many applications are emerging, both in academia and at the industrial level. A more complete description of the Semantic Web and its technologies is beyond the scope of this paper, and we will not go into details. The interested reader can find details in the vast literature which exists on this topic.

5.5 Why an ontological approach

The importance of semantic interoperability has been widely recognized by scholars, and many international projects agreed to use common metadata vocabularies (mainly based on Dublin Core metadata schema). This is a step forward towards the emphasis put in the last few years on XML data structuring. Scholars realized that XML is semantically poor, while the Semantic Web stack higher level technologies (RDF, OWL, etc.) can supply the appropriate technical environment to represent, export and share the knowledge needed to implement intelligent retrieval and browsing systems, and reason upon data. In the peer-to-peer web architecture, Semantic Web technologies permit fully decentralized semantic markup of content (for example, using classes and properties defined in CIDOC-CRM), and intelligent software agents can then use knowledge expressed by the markup.

In fact, looking back to the history of data cataloguing and sharing of cultural heritage information, we can see how we progressed from initial stages, where info was entered in an informal way, to more structured organization of information, and now we have many projects referring to a common metadata set (mainly Dublin Core, sometimes Qualified Dublin Core). Some more advanced projects (Hyvönen *et al.* 2004) rely on ontologies, mainly as a set of related terms to use for more precise queries. The question now, looking at the common agreement upon the metadata set, is *why should we consider an ontological approach*? First of all, as pointed out by Doerr (2003), even if both a core ontology and core metadata, such as Dublin Core, are intended for information integration, they differ in the relative importance of human understandability. Metadata is, in general, thought for human processing,

while a core ontology is a formal model for automated tools that integrate source data and perform a variety of functions. Vocabularies based on ontologies that organize the terms in a form that has clear and explicit semantics can be reasoned over, which is a fundamental process in enriching knowledge, inferring new information about resources. Secondly, there is a drawback in the implicit assumption made with the metadata approach. In short, it should become evident how adding metadata to the description of an artefact implicitly means that we assume a one-to-many (or possibly many-to-many) relationship between the object and the items identified by the metadata. Taking an example from art history, when specifying[8] some DC metadata like:

```
dc:title=Pietà
dc:creator=Michelangelo
dc:date=1499
dc:subject=Madonna
dc:subject=Christ
```

or

```
dc:title=Madonna del cardellino
dc:creator=Raffaello
dc:date=1505
dc:subject=Madonna
dc:subject=Child
```

we intend to say that a particular artefact (the Pietà, for example) was made by Michelangelo, is dated 1499, and has as its subject "Madonna" and "Christ", while the second one (the painting) was made by Raffaello, is dated 1505, and has as its subject "Madonna" and "Child". We can add controlled vocabularies to be sure that we specify correct terms for "creator" or "subject", but only humans can:

– check the consistency between dc:creator and dc:date as no artefact can be made by an artist after her/his death, or before her/his birth date (plus, let us say, 10 years?);
– having found an artefact, search for artefacts made in the same period, or by artists who were living and active in the same period;
– find the historical or political context (what was happening around these years?);
– find artefacts (for example portraits) which are "imaginary" portraits, because the scene is imaginary, or subjects never existed because they are mythological, or subjects did not exist at the time the artist was living or at the same time themselves.

[8] In the following examples, for the sake of simplicity we are not conforming to an actual syntax, which would require expressions like:
<meta name="dc.creator" content="Michelangelo" /> or
<dc:creator>Michelangelo</dc:creator>

It's worthwhile to recall how available thesauri are supposed to support some knowledge representation needs, but cannot be automatically translated into ontologies, as they sometimes model a class-subclass relationship (like "statues" and "korai (statues)"), sometimes model just different instances (for example, "Renaissance" is often modelled as a BT of "15th century", while both are periods in time, having some duration). Multiple inheritance and time dependent relationships are also an issue.

5.6 *Novelties and legacy*

The (Semantic) Web is opening new, fascinating scenarios, as an immense knowledge repository. Much information is conveyed by the links connecting different pieces of information. Web searching and browsing can take advantage of the interoperable knowledge representation to appropriately link information following the user's preferred interaction metaphors (spatial, temporal, classification affinity), thus greatly improving the access to information and knowledge stored in cultural web sites. In the Semantic Web environment intelligent user agents can rely on a core ontology to understand the mental model expressing the user's interests, implementing suitable navigation mechanisms (Signore 1995).

We can imagine (Signore 2004, 2005, 2006) an architecture where intelligent user agents can have access to the mental model expressing the interests of the user. The content can be tagged and semantically annotated using classes and properties defined in CIDOC-CRM. The agent can then perform reasoning, following the relevant associations and linking the information the user is interested in. The user's mental model can be expressed in terms of preferred interaction metaphors. Making reference to the ontology used as a basis for semantic annotation, this means specifying the set of classes and properties the user might be interested in navigating. Making reference to CIDOC-CRM classes, a user interested in the temporal context will be interested in classes like: E2.Temporal_Entity, E52.Time-span and their subclasses, at various levels, like E3 Condition State, E4.Period, E5.Event. The context can be expressed in a more precise way stating the properties the user is interested in (e.g. P117.occurs during, P118.overlaps in time with, etc.) to build up the temporal interaction metaphor. Identifying such properties can guide the agent to select the appropriate associations and perform the reasoning. The user agent (the browser) can be enhanced by two components: a reasoner and a finder, which accomplish the tasks of getting the semantic annotation of the current resource, looking to the user model, finding correspondences between user model and resource metadata, initiating a search following the properties the user is interested in.

We must stress, however, that the Semantic Web is just supplying the environment and technologies: ontologies, that play a central role in the architecture, must be filled in, otherwise the Semantic Web will never be

alive. Fortunately, there are decades, if not centuries, of studies that have built up knowledge. The problem is just to represent this knowledge in a more structured way, and make it sharable among different areas and usable by humans and machines. Therefore, all the past work is a precious legacy: scholars' knowledge must be formalized and made explicit as ontology, and very probably we will soon have to agree about a different model to represent objects, in a distributed and multicultural environment.

6. Conclusion

Cataloguing is an activity where knowledge plays a fundamental role. In the era of manual, paper based cataloguing, information was easily written down, but cataloguing cards were written by human experts, mainly for use by human experts, and the scholars' knowledge, which is the basis for putting objects into their cultural context, was essentially tacit, and unavailable. When computers started being used, it was necessary to store information in a more formalized way, and language normalization, data structuring, representation schemas came on to the scene. Cultural traditions lead to a description card approach, very similar to the one used in libraries. However, the intrinsic complexity of art and archaeological objects and their complex relationships required a thorough rethinking of the approach. In Italy we applied database conceptual design methodologies, ending in an object model that was a good trade-off between research and administrative needs, where information was fragmented in many small, semantically well defined, chunks. However, the main drawback remained the central role played by the object itself, with many attributes "predicating" its properties. As a consequence, the model was unable to represent the large variety of different semantic relationships among objects and, more important, with other pieces of knowledge pertaining to different disciplines.

The explosion of the web changed the traditional means of accessing information and focused attention on the interoperability issues. When they dealt with the problem of interoperability among different data sources, scholars realized that it was necessary to agree at least on a metadata based approach, such the Dublin Core common metadata set. However, the Dublin Core approach remains centred on objects, and information from different objects or disciplines can be merged only by a manual, human intervention, getting different records and filtering them appropriately.

An ontology based approach allows us to represent and share knowledge, and intelligent agents can infer new knowledge by automated reasoning on data. This is the Semantic Web scenario, where knowledge is available on multiple sources distributed over the entire world. The challenge is to represent, export and share expert knowledge which is the result of decades of studies.

It is not an easy task, but it is the way to achieve the goal of making expert knowledge available to any user, who would in this way be «able to search the online universe seamlessly as if the images and text about culture were available in one vast library of information» (Fink 1997).

Oreste Signore
Istituto di Scienza e Tecnologie
dell'Informazione "A. Faedo"
CNR – Pisa

Acknowledgements

First of all, an emotional memory of Roberto Gagliardi, a colleague of mine who was one of the most active and effective in the definition of the object model, and of Oreste Ferrari. He was both a scholar and ICCD Director, and his broad vision of the catalogue and of the mission of ICCD was essential to define objectives of high cultural level. I will never forget him as a gentleman with a subtle sense of humour. I wish to acknowledge people from ICCD, mainly Serenita Papaldo and Maria Ruggeri, remembering many years of joint work, when we developed and refined the object model, for their challenging requirements. Finally I wish to thank Irene Buonazia, who gave me some more details about the birth of cataloguing and mainly the ideas of Venturi and Argan.

REFERENCES

Baker T. 2000, *A grammar of Dublin Core*, «D-Lib Magazine», 6, 10 (http://www.dlib.org/dlib/october00/baker/10baker.html, accessed 9 June 2009).

Coyle K. 2007, *Managing Technology. The Library Catalog in a 2.0 World*, «The Journal of Academic Librarianship», 33, 2, 289-291 (preprint at: http://www.kcoyle.net/jal_33_2.html).

D'Amadio M., Simeoni P.E. (eds.) 1989, *Strutturazione dei dati delle schede di Catalogo. Oggetti di interesse demo-antropologico*, Roma, ICCD-Museo Nazionale delle Arti e Tradizioni Popolari.

Digicult 2003, *Towards a Semantic Web for Heritage Resources*, Thematic Issue 3 (http://www.digicult.info/downloads/ti3_high.pdf).

Doerr M., Hunter J., Lagoze C. 2003, *Towards a core ontology for information integration*, «Journal of Digital Information», 4, 1, n. 169 (http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/).

Fink E. 1997, *Sharing cultural entitlements in the digital age: are we building a garden of Eden or a patch of weeds?*, in *Museums and the Web: An International Conference (Los Angeles, CA, 1997)* (http://www.archimuse.com/mw97/speak/fink.htm).

Hyvönen E., Junnila M., Kettula S., Mäkelä E., Saarela S., Salminen M., Syreeni A., Valo A., Viljanen K. 2004, *Finnish Museums on the Semantic Web. User's Perspective on MuseumFinland*, in *Proceedings of Museums and the Web 2004 (Arlington, Virginia/Washington DC)* (http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html).

Massari S., Prosperi Valenti Rodinò S., Papaldo S., Signore O. (eds.) 1988, *Strutturazione dei dati delle schede di catalogo - Beni mobili storico-artistici: Stampe*, Roma-Pisa, ICCD-CNUCE.

Papaldo S., Ruggeri Giove M., Gagliardi R., Matteucci D.R., Romano G.A., Signore O. 1986, *Strutturazione dei dati delle schede di catalogo: Beni mobili*, in S. Papaldo, G. Zuretti Angle (eds.), *Atti del Convegno sull'Automazione dei dati del Catalogo dei Beni Culturali (Roma 1985)*, Roma, MiBAC-ICCD, 39-42.

Papaldo S., Ruggeri Giove M., Gagliardi R., Matteucci D.R., Romano G.A., Signore O. (eds.) 1988, *Proposta di strutturazione dei dati del catalogo: Beni mobili archeologici e storico-artistici (Edizione riveduta e aggiornata)*, Roma-Pisa, ICCD-CNUCE.

Papaldo S., Signore O. 1989 (eds.), *Un approccio metodologico per la realizzazione di una banca dati storico-geografica (A methodological approach to producing a historical/geographical databank)*, Roma, Multigrafica Editrice.

Parise Badoni F., Ruggeri M. (eds.) 1988, *Strutturazione dei dati delle schede di catalogo: Beni archeologici immobili e territoriali*, Roma-Pisa, ICCD-CNUCE.

Sheth A., Ramakrishnan C. 2003, *Semantic (Web) technology in action: ontology driven information systems for search, integration and analysis*, in U. Dayal, H. Kuno, K. Wilkinson (eds.), *Special Issue on Making the Semantic Web Real*, «IEEE Data Engineering Bulletin», 26, 40-48.

Signore O. 1986, *Architettura di sistemi per la gestione dei dati catalografici*, in S. Papaldo, G. Zuretti Angle (eds.), *Atti del Convegno sull'Automazione dei dati del Catalogo dei Beni Culturali (Roma 1985)*, Roma, MiBAC-ICCD, 51-58.

Signore O. 1993, *Cataloguing art objects: a comparison between French and Italian standards*, in D.A. Roberts (ed.), *European Museum Documentation Strategies and Standards. Proceedings of an International Conference (Canterbury 1991)*, Cambridge, MDA, 138-143.

Signore O. 1994, *From data structuring to data exchange: a simple path*, in H.J. Marker, K. Pagh (eds.), *Yesterday. Proceedings from the 6th International Conference Art History and Computing (Odense 1991)*, Odense, Odense University Press, 48-54.

Signore O. 1995, *Issues on hypertext design*, in *DEXA '95. Database and Expert Systems Application. Proceedings of the International Conference (London 1995)*, Lecture Notes in Computer Science n. 978, Springer Verlag, 283-292.

Signore O. 2004, *Representing knowledge in semantic cultural web*, in *EVA 2004 Jerusalem Conference on the Digitisation of Cultural Heritage (Jerusalem 2004)* (http://www.w3c.it/talks/eva2004Jerusalem/).

Signore O. 2005, *Ontology driven access to Museum Information*, in *CIDOC 2005 Documentation & Users. Proceedings of the CIDOC Annual Conference (Zagreb 2005)* (document: http://www.w3c.it/papers/cidoc2005.pdf; slides: http://www.w3c.it/talks/2005/cidoc2005/).

Signore O. 2006, *The Semantic Web and cultural heritage: ontologies and technologies help in accessing Museum information*, in *Information Technology for the Virtual Museum (Sønderborg 2006)* (http://www.weblab.isti.cnr.it/talks/2006/ITVM2006/).

ABSTRACT

Knowledge has been the driving force behind the Italian National Catalogue of Cultural Heritage. In the first stage, when the catalogue was mainly based on hand written paper cards describing objects regardless of their complexity, and intended for manual access by humans, the expert's tacit knowledge remained unexpressed, and the card had a simple structure.

Computer based applications initially relied on the features of Information Retrieval Systems, and simply converted typewritten cards into electronic documents. As results were quite disappointing, it became evident that a more formal representation of information was needed. The Italian experience led to the definition of a model for objects (simple, complex, aggregation of objects) with quite a large number of fields. Even if the schema was often perceived as too rigid, it proved to be effective for data exchange, and long lasting (the present XML model is almost the same, just with a different syntax). However, its main drawback was the "object centred" approach, and the impossibility of representing significant semantic associations with other disciplines. In this sense, a major objective, the contextualization of objects, remained unattained.

The web has been a "cultural revolution", because information is available everywhere, and users feel the need to combine different sources of knowledge. This semantic interoperability issue is often dealt with by adopting a metadata based approach (Dublin Core is the most popular). However, the metadata approach has the intrinsic limit that metadata are properties we "predicate" about items they refer to, and it is difficult, if not impossible, to derive new knowledge from the old. The Semantic Web perspective is much more ambitious, as the aim is to represent, export and share knowledge in a "machine understandable" way, and to allow intelligent agents to reason about it. In this light, scholars' knowledge must be formalized and made explicit as ontology, and very probably we will have to agree on a different model to represent objects, in a distributed and multicultural environment. This is not the end of the traditional scholars' knowledge, but a more effective environment for making this knowledge available to all users.