Semantic Web techniques for multimedia museum information handling

A. Stevenson[1], M. Addis[1], M. Boniface[1], S. Goodall[2], P. Grimwood[1], S. Kim[2], P. Lewis[2], K. Martinez[2],

[1] [mja,mjb,pg,as]@it-innovation.soton.ac.uk. IT Innovation, University of Southampton, SO16 7NP, UK.
[2] [sg02r,sk,phl,km]@ecs.soton.ac.uk. Department of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK.

Abstract: This paper describes the design and prototype implementation of a novel architecture for integrated concept, metadata and content based browsing and retrieval of museum information. The work is part of a European project involving several major galleries and the aim is to provide more versatile access to digital collections of museum artefacts, including 2-D images, 3-D models and other multimedia representations. An ontology for the museum domain, based on the CIDOC Conceptual Reference Model, is being developed as a semantic layer with references to the digital collection as instance information. The challenges of dealing with the import and management of multimedia museum information through the CIDOC CRM based semantic layer include data aggregation/derivation, cleaning, mapping and augmentation. Knowledge engineering precepts are used to guide the development of tools for data mapping and the organisation of workshops to enhance the museum partners' understanding of the ontology in relation to their legacy data. Augmentation agents for automatically extracting missing metadata instances from the Web are being developed as a way of assisting users to populate the semantic layer. Natural language processing techniques coupled with information retrieval tools are used for identifying information sources and structuring them into well-defined formats to be entered in the ontology. An ontological approach is also taken to the design and implementation of the user-interface of which a graphical concept browser is an integral component. This allows navigation through the semantic layer, display of thumbnails, or full representations of artefacts and textual information in appropriate viewers and the invocation of conventional content based searching or combined querying. Semantic Web technologies are used in system integration to describe how tools for analysis and visualisation can be applied to different data types and sources. This supports flexible and managed formulation, execution and interpretation of the results of distributed multimedia queries. Searches integrating concepts, content and metadata can be initiated from a single user interface. Combining system and data ontologies to define and publish the semantics of data retrieval brings closer the semantic web ideal of a self-describing digital heritage archive, accessible to anyone who can read these published semantics.

Introduction

The SCULPTEUR project takes an ontological approach to the semantics of search and retrieval, and to system integration. Five major European galleries are involved in the project: the Uffizi in Florence, the National Gallery and the Victoria and Albert Museum in London, the Musee de Cherbourg, and the Centre de Recherche et de Restauration des Musées de France (C2RMF) which is the Louvre related restoration centre. Each of these institutions maintains large digital archives of their collections. The textual and multimedia information contained in these archives, and in hundreds of other digital libraries and archives around the world, is diverse in both type and organisation. It represents an extremely valuable cultural heritage resource but despite the clear need for researchers, curators and historians to contribute to, develop and exploit these archives the ability of the these professionals to access the data is often limited. This is due both to the distributed nature of the resources and the difficulty inherent in trying search across heterogeneous systems. It is important to make the data archives available electronically, and to make then available in a structured way which encourages the development of knowledge built upon that raw data. Semantic Web technologies which can be used to generate, structure and link data can be applied to aid the development of that knowledge.

Cultural heritage institutions to need to develop, manage, visualize, navigate, search and exploit their valuable digital resources. The SCULPTEUR project [1], supported by the European Union under the Fifth Framework Programme, aims to fulfil these needs by building on the achievements of the ARTISTE [2] project and developing a system for integrated navigation and searching of gallery and museum collections using textual metadata, content-based analysis and ontological classification. This paper presents the design and prototype implementation of SCULPTEUR.

Motivation

Museums and galleries often own several different digital representations of some, or all, of the hundreds of thousands of works of art in their trust. These representations include public access images, specialized high-resolution scientific images used for conservation purposes, 3D models of individual artefacts and short movies showing artefacts in their gallery location. Data held in the collections owned by one gallery or museum is frequently relevant to the work performed at other galleries or museums. For example, when a museum conducts a programme of conservation and restoration it is beneficial to access information on the condition and treatments of similar works of art in other museums and galleries. However, discovery and access to this information is currently a manual and time-consuming process.

Different types of representation and digital textual metadata are often stored in separate collections and legacy systems. The heterogeneity of information and systems for museum collections creates several challenges for any system designed to support search and retrieval across that information, especially if the system needs to provide such services to other organisations or the general public.

The first, and most fundamental challenge is that the metadata terms used to describe and structure collections often differ from institution to institution. Previous approaches to this problem have focused on either imposing a common structure on the metadata to create a standard across a small number of institutions (Van Eyck project [3]), or imposing a standard interface technique (z39.50 [4], AQUARELLE [5]). However, these approaches fail to accommodate the diversity of specialized collections in the museum domain and fail to provide a schema that is sufficiently descriptive of the relationships therein.

The second challenge is to provide access in a way that enables the searcher to fully exploit the richness of the data available. Many existing digital library systems have a single entity, the digital text, at the centre of all user interactions with the system. The users search by specifying a word or phrase for the items of associated metadata, for example 'date of publication'. Similarly, previous multimedia digital libraries and image retrieval systems [28], such as ARTISTE, place the digital image at the centre of all user interactions with the system. The users search by specifying a value for one of the items of metadata associated with the image, for example "Find all images of art works where the artist's name is Raphael", or by specifying a content-based search in which case the user supplies a query image and asks to see, for example, "images of a similar colour".

A singular focus to the search specification and the objects returned does not allow the diversity of information often contained in a multimedia digital library to be fully exploited. For example, a typical museum library will contain metadata about art works (e.g. title, medium, state of restoration), the creators of those art works (e.g. name, date of birth), and digital representations of the art as well as metadata about those representations (e.g. angle of lighting, full or sub image), information about locations (e.g. as the place the art work is stored, or as the artist's country of origin) and dates (e.g. as the date of artist's death, or the date of creation digital image).

If all access is through the single concept of the digital image, then the user is forced to follow a single path through the data that obscures the multi-directional relationships that exist therein. However, access to the full complexity of the information in a user's collection, or in other collections, will only be useful if the complexity can be presented and navigated in a manageable way. As a result, graphical tools are required for ontological browsing of the concepts, relationships and instances within collections.

A third challenge arises when using content-based search techniques [20,23,30] such as searching the colour, texture, shape etc of the digital representations. There is significant value to the user when content-based techniques are combined with textual metadata searching [19,21,24,25]. However, content-based analysis and comparison techniques are highly specialised, often computationally intensive, and have little or no support for describing their semantics in current tools and standards. Furthermore, it is unreasonable to expect all museums and libraries to support a common set of content-based search techniques. Therefore, if a user seeks to search a collection they need to first determine the search capabilities supported and then formulate an appropriate query. Whilst this is manageable for occasional searches of single collections, there are clearly problems when looking to automate distributed queries across multiple collections. Of course, the use of a common query language syntax and a common protocol for search and retrieval is not precluded, and indeed these can be used to provide the backbone of interoperability.

Finally, information about museum collections is sometimes incomplete or may even contain inaccuracies. Semantic Web [8] and Agent technologies offer a way to find additional information on the Web. This additional information may be missing items of metadata associated with an artefact or artist, or may be additional narrative and supplementary information of interest to end-users such as artist biographies or further sources of information available on remote Web Sites. Augmenting museum collections is not limited to adding information from external sources. Existing content can be analysed to further classify the items in the collection, for example classifier agents could monitor the collection and use known associations between content descriptors (colour, texture, shape, volume) to aid classification of existing objects or new acquisitions.

Overall, there is a clear need for galleries and museums to be able to better augment, navigate, exploit and share the rich information in their digital collections. Galleries and museums will wish to maintain autonomy over the way their content is structured and they will want control over what services are provide to their users, both internal and external. This requires SCULPTEUR to abstract the complexity and heterogeneity of their legacy systems to provide simple browsing and search facilities for the user using a combination of ontological, textual metadata and content-based analysis techniques. Furthermore, published semantics and interoperability protocols are needed to allow individual SCULPTEUR systems to interoperate to achieve seamless cross-collection searching.

Architecture

The SCULPTEUR architecture is designed to provide integrated concept, metadata and content based browsing, retrieval and analysis of museum information. The architecture includes components for augmenting the knowledge about the museum's collections through semi-automated classification of its content and through information extraction from external sources such as the Web.

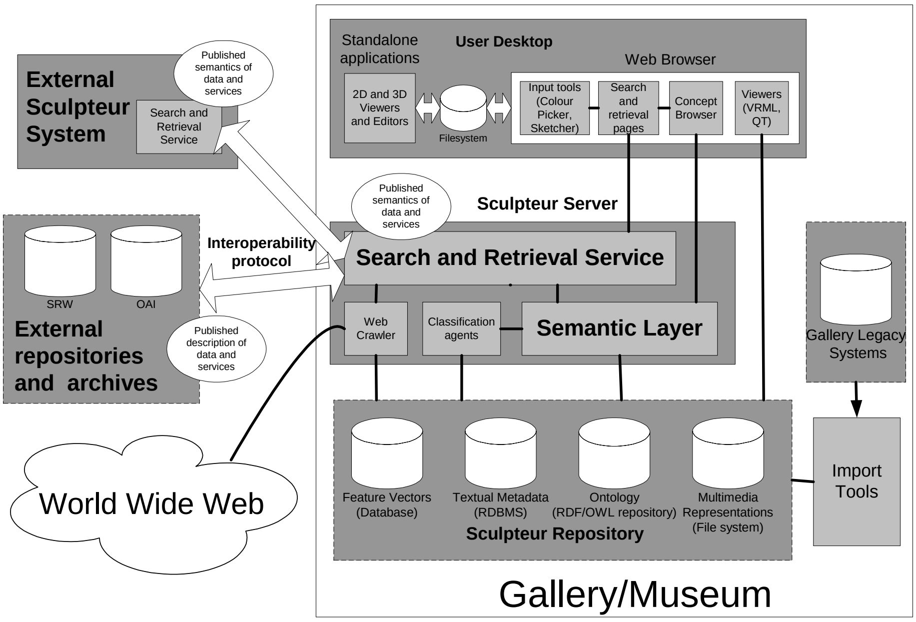

Figure 1 Sculpteur Architecture

As Figure 1 illustrates, SCULPTEUR will include tools for importing data (new images, 3-D models and other digital representations of museum artefacts, data from other gallery systems and information from the web) and for interoperating with external systems (remote SCULPTEUR installations, gallery legacy systems or remote digital libraries supporting standard interfaces such as OAI[6]).

Further details of the components of the SCULPTEUR architecture can be found below.

The Semantic Layer

To solve the problem of a singular focus obscuring potential interesting information to be derived from links between data, SCULPTEUR will employ a semantic layer that makes explicit those entities and relationships which are implicit in the data. Clearly the information that a particular artist was born in a particular country is not metadata about a digital image, it is metadata about that artist. The semantic layer means that users can make any of the entities identified in the collection as the focus of their search, both in terms of the search specification and the objects returned by the search. Example searches might include “Find all countries that have produced artists working in the 19th Century”, “Find the average size of art works made of stone in America”, or “Find images showing the back of art works where those art works have been restored". In the same way that there is no centre of the Web there is no centre of the SCULPTEUR system – while one user may be interested in searching for artists, another may be interested in place and yet another in the art works. The SCULPTEUR system will enable users to approach the same data in multiple different ways using a single interface.

The semantic layer consists of the ontology [9] and instance information. The associations between object representations and concepts in the ontology allows enhanced retrieval possibilities by facilitating broadening or narrowing of the search scope through broader and narrower concepts. It will also permit enhanced content based searches since a content match with an object associated with a concept will allow other objects associated with the concept to be retrieved whether or not they are visually similar. However the choice of ontology and the question of how concepts in that ontology map to the instance information raise several interesting issues.

Ontology

Legacy metadata tends to use domain-specific and often museum specific terms making it difficult to search multiple collections. Standard metadata vocabularies such as the Dublin Core metadata standard[29], Iconclass[35] or the AAT[36] make it possible to establish equivalencies between collections and they therefore fulfil the information integration requirement upon any candidate ontology. The Dublin Core was used for this purpose in the ARTISTE project. However to create the rich knowledge space which we desire, something more complex which can model all the entities and their relationships present in the instance information is required. To build our own ontology would be a labour-intensive and time-consuming task since it is necessary to make sure the hierarchies of the concepts are structured accurately and without duplications. For this reason, and because, in the pursuit if interoperability, it is always preferable to use existing standards where possible we examined the CIDOC CRM.

The CIDOC Conceptual Reference Model (CRM) developed by CIDOC Documentation Standards Working Group is concerned with cultural heritage information describing concepts and relations relevant to all types of material collected and displayed by museums [10]. It aims to support the exchange of relevant information across museums through coherent semantics and common vocabularies. The CRM is not the only publicly available ontology which can be used to structure cultural heritage information as the work done by the Harmony project to map metadata from CIMI members to the ABC demonstrates [33]. However users of the CIDOC CRM represent a growing community as proven by the OntoWeb Project [39] which is investigating technologies and contents for the Semantic Web related information exchange processes and has included the CIDOC CRM in the list of promising standards to support information exchange in the Semantic Web. Other standards communities are also recognizing CIDOC CRM as a valid standard and attempting to integrate it into different areas. For example one project has undertaken to express the ontology using DAML + OIL [37] and Jane Hunter's work on merging the CIDOC CRM and MPEG 7 proposes using the CIDOC CRM as the underlying ontology and extending it to provide additional multimedia/MPEG-7-specific classes and properties [38]. The ABC ontology is not so widely recognised or used and we therefore decided to base the prototype implementation of the semantic layer upon the CRM.

SCULPTEUR has extended the CRM by adding concepts and relationships such as "painting" and "has_related_artists" (a relation linking to other artists involved in the creation of the painting) and by including multimedia representations as artefact instances. shows part of the extension of CRM in SCULPTEUR. Protégé (a graphical tool for ontology creation and maintenance) was used for such modifications [11]. The underlying expression of the ontology is in RDF/RDFS.

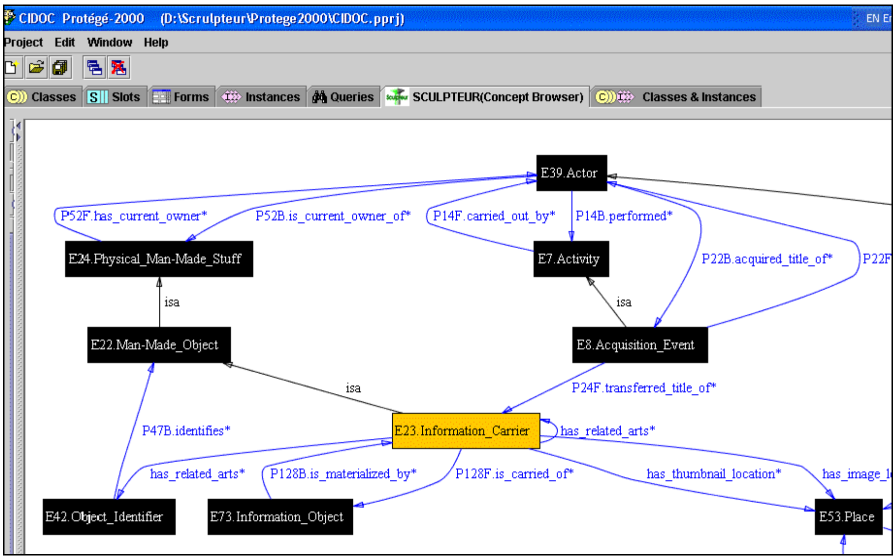

Figure 2 Protégé screen shot showing an example of extensions to the CRM.

Ontology Mapping

The question of how the ontologies used in the semantic layer map to the instance information introduces the possibility of various refinements to the architecture. The integration of the legacy museum information with the semantic layer can be achieved either by the storage of instance information within a knowledge base, or through mapping concepts in the ontology directly to attributes of the data source. An example of the latter is the mapping of the 'author' concept in the ontology to the 'createdBy' column of a particular database table.

Mapping between ontologies and data sources can take either a centralised or distributed approach. In a centralised architecture, all data from museum and gallery legacy systems are mirrored in a central location, and a mapping between the ontology and the central data source is created. A user can query the system or other SCULPTEUR knowledge bases using a uniform interface. With a distributed knowledge base, data is not copied to a central location but remains within existing legacy systems. These legacy systems are not limited to other SCULPTEUR knowledge bases, but could include data sources such as OAI repositories and SRW [7] services as well. Each data source that forms part of the knowledge base would have to be mapped to the SCULPTEUR ontology. Using a distributed knowledge base would increase the complexity of query execution. Users would compose queries in a single user interface that would be represented in the system as a common query language. This common query language would then have to be translated to queries that run on specific data sources such as OAI or SRW.

There are positive and negative qualities of each of these approaches. Simplicity characterises the centralised approach, regardless of whether information is stored in a relational database or as an RDF [32] file. Storage in RDF is likely to be significantly slower, however. The negative side to the centralised approach is that searching is constrained to SCULPTEUR knowledge bases, and in addition there will be data consistency issues. A distributed approach avoids the consistency problems and the limitations on the search arena, but introduces new problems associated with a more complex query execution process

Whichever architectural approach is taken the formation of the mappings between the ontology and the instance data still remains a not inconsiderable task. The process of mapping between data different data models (e.g. from a legacy relational database schema to the CIDOC CRM) has many issues including data cleanliness, consistency, completeness and losses. Typically, data has to be pre-processed to address these issues if, when imported into another system, it is to produce sensible results. In addition, a detailed understanding of the schema, usage and record values is also required so that correct mappings can be created.

For example the metadata to be imported into the SCULPTEUR prototype included about 7000 art object records and 47,000 object representation records and presented several problems:

The fields that were defined as “controlled” contained inconsistent values. The allowable values in the controlled lists are defined within museum documentation but not enforced within the application or database leaving it up to the user to ensure that they are typed correctly. For example, a controlled list is defined for view of object that should contain values such as “top”, “bottom” etc. However, in the metadata there are values such as “top;” “toop”. If the column were to be mapped to a concept in the CRM for the view of the object then each new spelling of “top” would be a different view.

Many of the fields in the database are empty which resulted in a sparse population of the CIDOC-CRM. Since the CRM is generally much richer than the existing database schema this will probably be a situation which reoccurs for most museums.

Several fields can contain similar information, but from different perspectives. A gallery collection can contain metadata for several different purposes (curation, conservation, photograph services etc.). This can result in several metadata fields that contain similar information, but from different perspectives. For example, an art object may be described from the point of view of collection management but also from the point of view of commercial photograph sales. Although the fields describe the same physical art object, the people authoring the content and the terminology they use can be quite different. Furthermore, neither viewpoints can, or should, necessarily be treated as definitive. This obviously makes mapping harder due to dealing with inconsistencies and ambiguity. This can be seen in Figure 3 below where there are similar fields describing the art object (from a collections management perspective) and the photographs of that art object (from a picture services perspective).

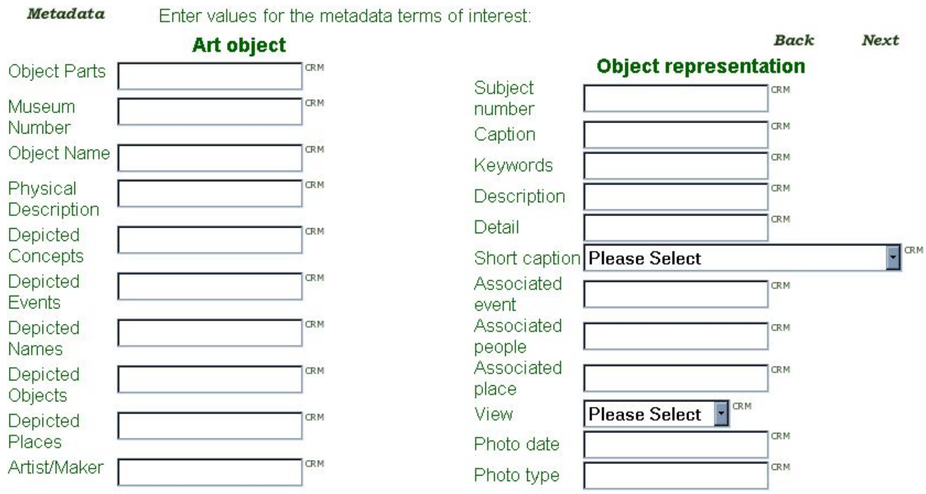

Figure 3 Example of fields containing similar information from different perspectives

Some simple metadata attributes describing an art object or representation can actually correspond to a long chain in the CRM. For example, a simple attribute like 'photograph date' has a long-winded mapping in the CRM:

```
E84 Information Carrier -> E77 Persistent Item ,
P92B was brought into existence by , E65 Creation
```

```
Event , P4F has time span , E52 Time Span , P82F at
some time within String
```

Some metadata fields contain complex/non-atomic values that express relationships between records. It was assumed that it would be possible to map table columns to concepts in the CIDOC-CRM, however, this is not always possible or sensible. For example, **Figure 4** shows a controlled list for the category field in the a legacy system. It can be seen that the field values actually contains a hierarchical thesaurus whose structure is defined by semi-colons. Mapping the Category field directly to a concept does not express the richness of information contained within this field. If the category classification could be extracted from this field into a thesaurus then the hierarchical relationships expressed in that thesaurus could be incorporated in the ontology.

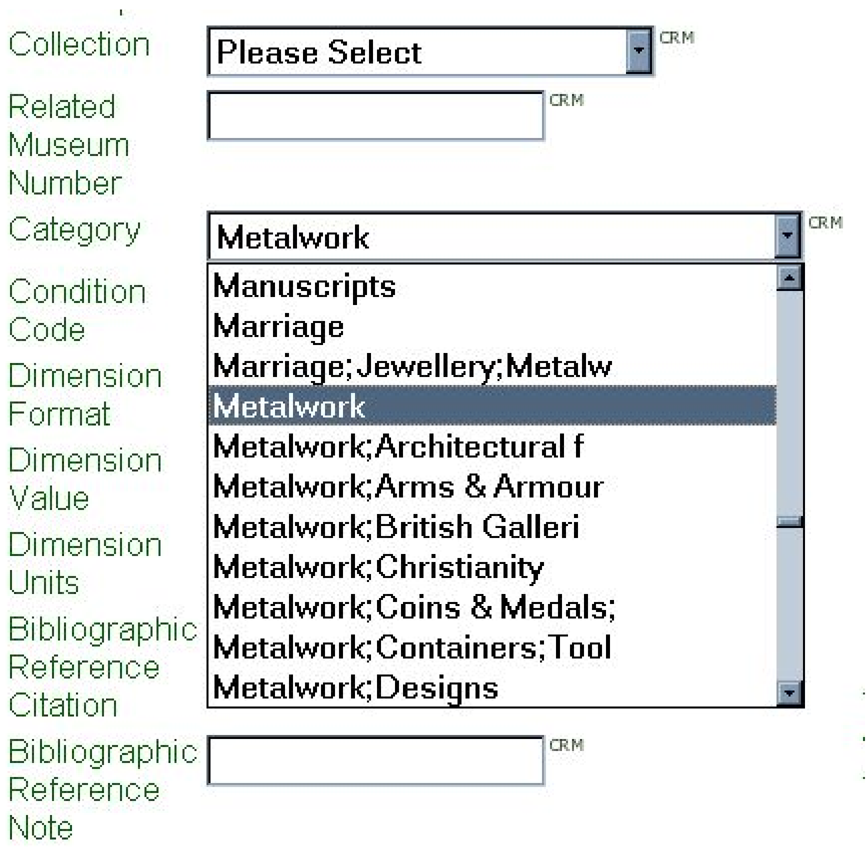

Figure 4: Complex field in data from a legacy data system

Descriptive metadata fields often contain free text that can contain information that relates to several parts of the CRM, e.g. people, places, object descriptions and events. This can be seen in Figure 5 below. These rich descriptive fields are most valuable when it comes to search and retrieval, but are also the hardest to map since they require extensive pre-processing.

Figure 5 Example of descriptive metadata fields

Clearly tasks like transforming controlled lists into well formed hierarchical thesauri and making the best mappings between relational database fields and ontology concepts requires the expert domain knowledge of the museum and gallery professionals. To support the SCULPTEUR museum partners in this task a workshop was held which followed knowledge engineering precepts, specifically the CommonKADS methodology. The museum partners were introduced to the CIDOC CRM and encouraged to begin the process of mapping their legacy database schema to the model and thus incorporating their data into the semantic layer.

Automatic Augmentation

One area in which the ontology itself can aid the population of the semantic layer is through augmentation agents. Missing instance information is common since galleries hold large numbers of artefacts and the relevant information is scattered in different places, which makes it difficult to collect and locate all values. Examples might be missing dates when works were created or names of places where artists were born or worked. In addition, the knowledge required for identifying and validating such knowledge is generally held by a small number of people. Hence, the development of tools for automatically finding missing instances is useful in order to reduce human efforts and time in processing related information. The Web, in particular, is potentially useful for gathering such missing information, but sites conforming to semantic web standards are limited. Information processing is required for extracting the missing relations from the Web. Whereas search engines (e.g. 'Google' or 'Yahoo') can retrieve pages which are possibly relevant to the information required, the extraction of the specific relations within the pages can be better served by knowledge extraction techniques. We are developing searching "agents" which can identify and, where possible, retrieve missing information automatically. GATE [12] and WordNet [13] are applied to retrieved web pages in order to extract named-entities for identifying and completing the missing instances. Since these relations generally link two concepts conforming to ontology specifications, it is necessary to construct any semantic connections between identified entities. A syntactic and semantic analysis, based on a natural language processing, can structure the extracted annotations according to their roles in a given sentence (e.g. 'subject' or 'object'). Based on subject-verb or object-verb associations, a set of potential relations are created. Human experts are used to validate correctness of information before it is committed to the knowledgebase.

The Concept Browser - an interface for combined semantic and content-based queries and browsing

While some of advantages of the semantic layer have already been enumerated above we also believe that the ability to navigate a collection using the concepts in the ontology will be a boon for users, particularly those without a detailed knowledge of the collection who might find it difficult to specify a traditional query.

However exposing the complexity of the CIDOC CRM to users presents its own problems. As Doerr, Hunter and Lagoze point out it is the nature of an ontology which seeks to provide a underlying formal model for complex data integration and the exposure of multi-directional relationships, that its "design should be motivated more by completeness and logical correctness than human comprehension" [34].

Therefore, as shown in Figure 6, we have developed a Concept Browser to allow graphical navigation directly through the ontology as a visualization of the semantic layer.

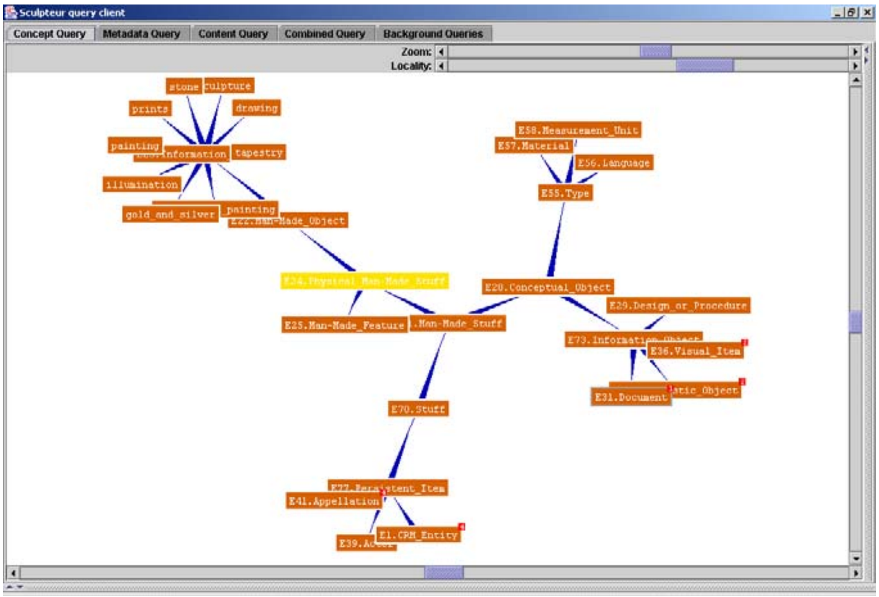

Figure 6 Concept Browser showing Extensions to CICOC CRM

A graphical concept browser provides user access to parts of the ontology allowing searchers to navigate easily to representations of interest. Combined with algorithms for multimedia object matching (based on feature vectors extracted from the representations) and menus for metadata selection, the system facilitates individual or combined concept, metadata and content-based searches.

One of the difficult HCI issues is how to allow people to form a query composed of a semantic term plus a content-based query in the form of a choice of image and algorithm. In this interface three different information areas are displayed as tabbed panels: concept, metadata (i.e. things which are not extracted into the concept layer) and image content queries. By selecting items from each panel a combined query can be composed such as “made of wood” and “like this image in terms of colour histogram”.

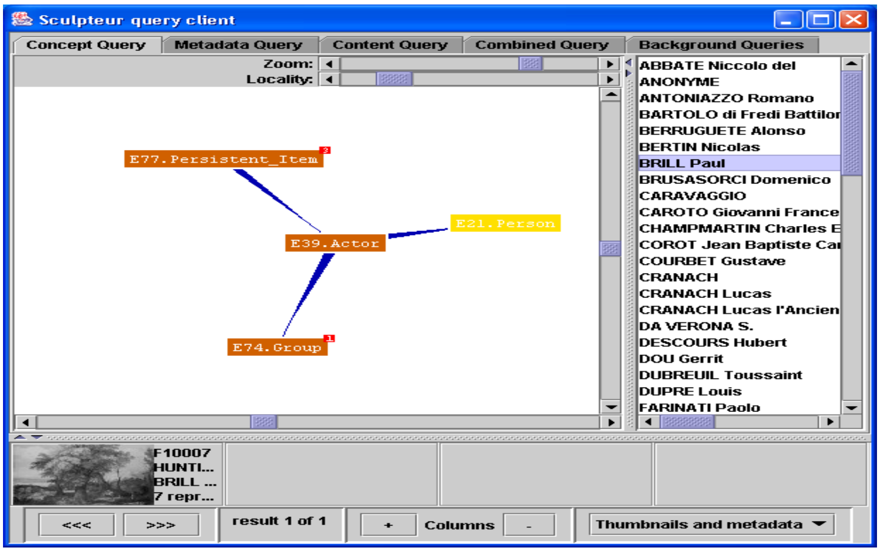

Figure 7 User Interface to the Concept Browser

Each panel can be used individually and the concept panel is particularly interesting as it gives a visual representation of the ontology. Clicking on a concept puts it into the centre and shows concepts close to it. Changing the "locality" allows more distant terms to be seen together. Right-clicking on a concept allows the user access to request all the instances, as shown in above. The Concept Browser proved to be a useful test interface and it was found that the approach was relatively intuitive. However, plenty of screen space is required and the location of concepts on the screen was not fixed which made "visual hunting" necessary.

In addition, since user preferences for browsing or navigating the ontological space may be considerably different from the ontology hierarchy based on the CRM, we are investigating various approaches for dynamically re-structuring the CRM hierarchy according to the user preferences. One of the methods of interest to us is to hide/remove the intermediate concepts and to link directly the start concept and the destination concept preserving the roles of the intermediate concepts.

Defining the Semantics of search and retrieval

It is important to model the semantics of the search and retrieval processes to support systems integration and to assist the user in the search and retrieval process.

Queries supported in SCULPTEUR are: **textual metadata** (existing descriptions of objects in a collection stored in relational database tables); **representation content** (feature vectors representing shape, colour, texture etc.); and **concepts** (things in the domain ontology, for example 'painting', 'sculpture', 'style'). For example, a user can retrieve museum objects according to associated textual metadata by filling in required values in a form. Content-based analysis can be used to return similar representations (2D images, 3D models, image movies) to one supplied by the user. Browsing or searching through concepts enables a user to locate and retrieve museum objects according to the concepts with which they are associated, or by relationships to other concepts.

These three apparently different areas can all be treated in the same way by considering them all to be ontological concepts, i.e. by building a unifying model of all aspects of the search and retrieval domain. In doing this all searches can be done in an integrated and uniform way. Single statements can be used to specify sophisticated queries such as 'find the 2-D thumbnails for all oil paintings that are authored by Van Gogh, where the painting contains colours similar to the oranges and yellows that I select'.

However the SCULPTEUR system needs to do more than simply integrate different types of query since SCULPTEUR incorporates a diverse range of tools for use in the formulation, execution and interpretation of the results of a query. This complexity is illustrated in Figure 8. For example, the inputs to a multimedia search and retrieval query can be supplied in a variety of ways such as query images or 3D models, textual metadata, concepts found by browsing an ontology, or free text. Similarly, the outputs could be images, models, concepts, or textual information. Furthermore, it needs to be possible to view these outputs in many ways such as 2D thumbnails ordered by similarity to a query image, art object titles chronologically ordered on a timeline, or paintings presented in a virtual gallery. Other ways of presenting results could include locations on a museum floor plan, or a street map showing gallery locations. A range of tools will be required to support this diversity of inputs and outputs.

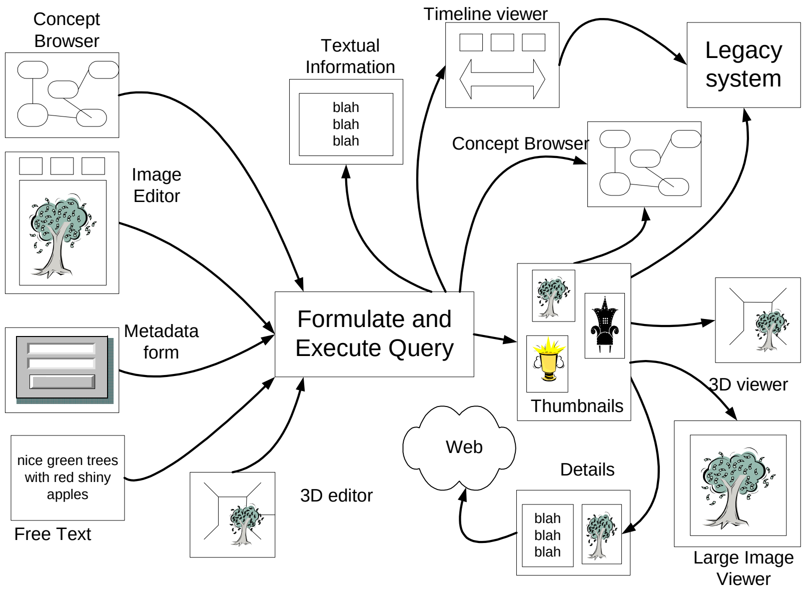

Figure 8 Inputs and outputs of the SCULPTEUR search and retrieval process

SCULPTEUR has a system ontology which includes concepts such as the digital representation of art objects (JPEG, TIFF, VRML, 3D models), feature vectors of those digital representations (colour histograms etc), algorithms used to produce and compare feature vectors, tools to construct queries (colour picker, 3D editor) and to display digital representations (2 and 3 D viewers), and other components of the search and retrieval process such as QueryRepresentation (2D or 3D representation supplied as input to a query) and ResultSet (things returned as results of the query).

We have prototyped the system ontology using Protégé [11]. We then developed queries to determine if the ontology could be used successfully to support the search and retrieval process. Some of the classes and relationships in the system ontology are presented in Figure 9. This diagram shows the high-level system concepts and their relationships. Not all of the ontology is shown, in particular the range of representation types and formats is limited, and the link to the museum domain is not shown. Specific relationships such as 'the Colour Coherence Vector algorithm can be applied to 2D colour images to produce a Colour Histogram'. These relationships are modelled through the idea of DigitalAttributes. A DigitalAttribute (e.g. has colour) captures the specific requirements of each Algorithm and also describes Representation types that are suitable for providing this input.

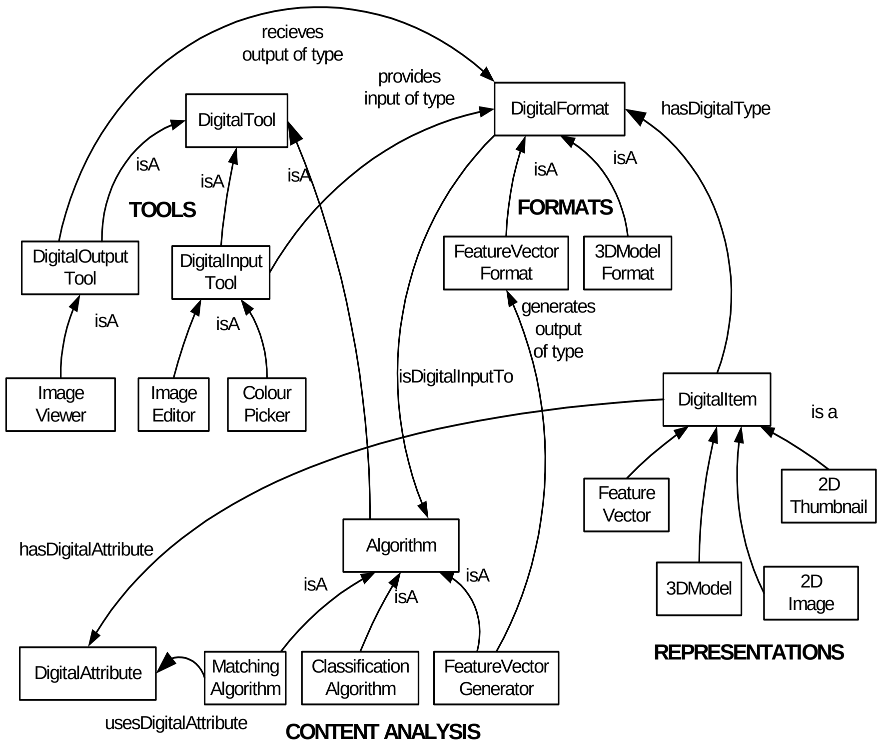

Figure 9 Extract from Sculpteur system ontology

SCULPTEUR will use the system ontology to determine the right tools, data and algorithms to use for a particular query. The process of execution of a combined concept and content-based query such as “Find and then show me 3D models of vases that are similar to a 3D model of a vase in my collection” illustrates this idea. The system would interpret the users’ intentions and use the ontology to present the appropriate tools to the user: (i) a 3D viewer or editor and (ii) a means to specify the concept Vase. An appropriate algorithm for the comparison of feature vectors of 3D models would be selected by using relationships that link media formats and associated algorithms that can process them. On completion of the query the system would use the appropriate tool capable of displaying the 3D models. A further advantage of an ontology describing the SCULPTEUR system is that it enables new components to be added as required to the system (such as a new algorithm, or an additional VRML viewer) without extensive recoding of interfaces or application servers.

The process of a combined metadata and content-based search for the ‘Van Gogh painting’ example given earlier further illustrates how the system ontology and domain ontology can be combined to facilitate query composition and query execution. The query process can be broken down into a number of steps as outlined below. Each of these steps were developed and tested using Protégé against a combined system and domain ontology, and therefore some of the description below refers to concepts not shown in Figure 9. In the full system, the users will not be subject to this query process since it will be transparently and automatically executed behind a much more intuitive user interface.

- **Q1: Find all 'MatchingAlgorithm' that 'usesDigitalAttribute' Colour.** This would return the appropriate matching algorithm.
- **Q2: Find all 'DigitalFormat' that 'isDigitalInputTo' Q1.** This would return the ColourHistogramFormat instance of 'FeatureVectorFormat' concept. This is necessary information to find appropriate tools and to limit the feature vector space during query execution
- **Q3: Find all 'DigitalInputTool' that 'generatesDigitalFormat' Q2.** This would return instances of the ColourPicker concept. The ontology has been now been used to specify the correct tools available for use with a particular query.
- **Q4: Find all 'Painting' that 'wasCreatedBy' Van Gogh.** This is a metadata search.
- **Q5: Find all '2DImage' 'represents' Q4.** The system concerns itself with digital representations rather than original art objects so we need to find those that represent the paintings returned by Q4.
- **Q6: Find 'FeatureVector' that 'hasDigitalType' Q2 & 'FeatureVector' 'isFeatureOf' Q5**. The search space has been constrained to those feature vectors that are of the correct type and that represent 2D images of painting by Van Gogh.

The feasibility of using ontologies for these system aspects of SCULPTEUR depends strongly on the availability of suitable tools and standards. We believe that RDF and RDFS [32] are sufficiently expressive, and that both suitable tools and query languages [31] exist to allow SCULPTEUR to automatically query the system ontology for the information it needs.

As well as support for the facilitation of systems integration, the approach taken of defining aspects of the SCULPTEUR system within the ontology also provides important support for the search and retrieval process. Query execution in SCULPTEUR will often be an iterative process: A user performing an initial search may wish to then either broaden or narrow the set of returned results, or use one of the returned images in a subsequent query. This search process would flow until the correct item/s are found. Containing the semantics and provenance of aspects of the search and retrieval process - such as a QueryRepresentation or ResultSet - within the ontology provides for a powerful iterative search and retrieval process. Moreover, defining semantics of search and retrieval provides specific support for interoperability and cross-collection searching.

Interoperability

The extensive use of semantic web technologies in SCULPTEUR provides a way to establish common semantics between heterogeneous digital libraries containing multimedia collections. These common semantics include how to perform, content-based analysis and searching by concept as well as conventional textual metadata searching. This goes a long way towards interoperability between multiple digital libraries. In fact, this would be sufficient to enable cross-collection search and retrieval between SCULPTEUR systems. However, common semantics are not enough to provide interoperability with third-party systems. To achieve this requires adoption of standard protocols for the process of search and retrieval.

In order for the digital heritage resources contained within a museum or gallery collection to be incorporated into the semantic web, it is essential that both the digital resources themselves and the features of the system be made available in a machine understandable way. By describing aspects of the system and the search and retrieval process - such as the algorithms supported and the nature of the results of a search - as concepts in the ontology, the necessary semantics for interoperability are provided. In essence, the system has to describe what is available and how to use it.

The Search and Retrieve Web Service (SRW) [7] is an initiative based on the z39.50 protocol [4] for searching databases that contain metadata and objects. It proposes a query language, Common Query Language (CQL) [15] and currently represents the emerging standard in the area of distributed data access for digital libraries. However being traditionally concerned with text based searching, CQL provides no formal method to specify searches for similar images or 3D models using particular content-based algorithms [14]. Although many multimedia oriented query languages exist [16][17][18] for content-based searching, there is no one language that has been accepted as the standard multimedia query language and that addresses the issues related to searching distributed collections.

ARTISTE was one of the early implementers of SRW [27] and worked in close contact with the z39.50 community to develop the SRW specifications to extend the capabilities to combined image content and metadata based searches over multiple collections. CQL was expanded to provide support for image content queries by adding image operator (img-op), image analyser (img-analyser) and an image expression (img-exp) to the language.

```
primary ::=result-set-expression | [index-name rel-op]
adj-expr |  index-name img-op img-analyser img-exp
```

The SRW CQL specification of result-set-expression and index-name remains unchanged in the SCULPTEUR CQL. The SCULPTEUR CQL further specifies elements necessary to an image content query

```
img-op ::= “SimilarTo” | “PartOf”
```

```
img-analyser ::= identifier

img-expr ::= url
```

It can be seen from the definition of img-expr above that query images are specified as URLs. The same approach is used for query result images. Some examples of SCULPTEUR CQL queries are given below.

```
dc.Creator contains Vinci and
sculpteurCore.VisibleLightImage SimilarTo CCV
http://scutlpeur.it-
innovation.soton.ac.uk/test_images/test.jpg
```

This query combines a Dublin Core [29] ‘Creator’ metadata search with an image content-based query that uses the ‘CCV’ (Colour Coherence Vector) algorithm to find images that are ‘SimilarTo’ the referenced query image ‘test.jpg’.

```
dc.Subject = TEXTILE and
sculpteurCore.VisibleLightImage PartOf MCCV
http://sculpteur.it-
innovation.soton.ac.uk/test_images/test.jpg and
dc.Creator contains Morris and William
```

This query combines a textual metadata search involving the Dublin Core attributes ‘Subject’ and ‘Creator’ with an image content-based query that uses the ‘MCCV’ (Multiscalar Colour Coherence Vector [22]) algorithm to find images that have the referenced query image as ‘part of’ them, i.e. as a sub-image.

Conclusion

In this paper we have presented a new approach to the design and implementation of a system to provide effective searching, navigating and querying the diversity of multimedia information held by museums and galleries. A coherent architecture is achieved by using ontologies to describe both the domain of the stored information and the features and facilities of the system itself.

The design supports content, metadata and concept based approaches to querying and these can be used together in a seamless way or separately when required. The ontology for the museum domains can be navigated using a graphical concept browser, which also gives direct access to the multimedia representations. Agents are being developed to extract missing information using semantic web and natural language processing technologies.

While highlighting some of the real-world data integration problems faced by those attempting to use complex ontologies to structure legacy data we have also enumerated the many advantages to that approach. We have begun to show that it is possible to use a combination of system and data ontologies to define and publish the semantics of search and retrieval. In this way we are moving towards the semantic web ideal of a self-describing digital heritage archive, accessible to anyone who can read these published semantics.

Acknowledgements

The authors wish to thank the European Commission for support through the SCULPTEUR project under grant IST-2001-35372 and collaborators on the project for many useful discussions, use of data and valuable help and advice. The authors wish to thank P Dibdin, A Pillinger, R Sadotra, A Smithson and T Wirdyanto who, as Master of Engineering students, contributed substantially to the development of the SCULPTEUR prototype. Finally, we are grateful to TouchGraph (www.touchgraph.com) for software used in the concept browser and Hewlett Packard's Art & Science programme for the kind donation of server equipment.

References

[1] SCULPTEUR IST-2001-35372 http://www.sculpteurweb.org
[2] ARTISTE IST-1999-11978 http://www.artisteweb.org/
[3] J. H. E. van der Starre, "Visual Arts Network for the exchange of cultural knowledge: Van Eyck Project", Information Services & Use, Vol.13, pp 347-355, 1993.
[4] z39.50 http://lcweb.loc.gov/z3950/agency/
[5] A. Michard, V. Christophides, M. Scholl, M. Stapleton, D. Sutcliffe, A.M. Vercoustre, "The Aquarelle resource discovery system", Computer Networks and ISDN Systems, Vol. 30, 1185-1200, 1998.
[6] Open Archives Initiative http://www.openarchives.org/
[7] ZING Search and Retrieve Web service http://www.loc.gov/z3950/agency/zing/srw/
[8] The Semantic Web http://www.semanticweb.org/
[9] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993
[10] N. Crofts, I. Dionissiadou, M. Doerr, M. Stiff, Definition of the CIDOC Object-Oriented Conceptual Reference Model, V.3.1, July 2001
[11] J. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, S. W. Tu, The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. 2002. Technical Report, SMI-2002-0943
[12] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, US.,2002
[13] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An On-Line Lexical Database, In Technical Report, University of Princeton, U.S.A., 1993
[14] ARTISTE D6.2 Impact on World-Wide Metadata Standards http://www.it-innovation.soton.ac.uk/artiste/documentation/D6.2.final.pdf
[15] CQL http://www.loc.gov/z3950/agency/zing/cql/cql-syntax.html
[16] J.Z. Li, M. Tamer Ozsu, D Szafron, V. Oria. MOQL: A Multimedia Object Query language. The 3rd Intl. Workshop on Multimedia Information Systems, 19-28, 1997
[17] A.Henrich, G Robbert. POQL: A Query Language for structured multimedia documents. Proceedings 1st Intl Workshop on Multimedia Data and Document Engineering, 17-26, 2001

[18] G. Amato, G. Mainetto, P. Savino. A Query Language for Similarity-based Retrieval of Multimedia Data. Advances in Databases and Information Systems, 1-13, 1997.

[19] Addis, M., Lewis, P., Martinez, K. "ARTISTE image retrieval system puts European galleries in the picture", Cultivate Interactive http://www.cultivate-int.org/issue7/artiste/

[20] F. S. Abas and K. Martinez (2002) Craquelure Analysis for Content-Based Retrieval. IEEE DSP 2002 conference. July 2002.

[21] Dobie, Mark R. and Tansley, Robert H. and Joyce, Dan W. and Weal, Mark J. and Lewis, Paul H. and Hall, Wendy (1999) A Flexible Architecture for Content and Concept Based Multimedia Information Exploration. In Harper, David J. and Eakins, John P., Eds. Proceedings The Challenge of Image Retrieval, Newcastle, 1999, pages 1-12.On Fri, 2 May 2003,

[23] M. Addis, M. Boniface, S. Goodall, P. Grimwood, S. Kim,P. Lewis, K. Martinez, A. Stevenson, Integrated image content and metadata search and retrieval across multiple databases, To be presented at CIVR2003, July 24-25, 2003, University of Illinois at Urbana

[24] J Stevenson, A Review of the Artiste Project, International Conference on the Challenge of Image and Video Retrieval CIVR2002 July 19, 2002, London, UK http://www.civr2002.org

[25] P. Allen, M. Boniface, P. Lewis, K. Martinez "Interoperability between Multimedia Collections for Content & Metadata-Based Searching", 11th WWW Conference, Hawaii. 7-11 May 2002 http://www2002.org/CDROM/alternate/196/index.html

[27] ARTISTE SRW Service demonstration http://artiste.it-innovation.soton.ac.uk/srw

[28] R.C. Veltkamp and M. Tanase. Content-based image retrieval systems : A survey. Technical report UU-CS-2000-34, Department of Computing Science, Utrecht University, October 2000. 34

[29] Dublin Core `http://www.dublincore.org/`

[30] Arnold W. M. Smeulders, M Worring, S Santini, A Gupta, and R Jain. Content-Based Image Retrieval at the End of the Early Years. In Transactions On Pattern Analysis And Machine Intelligence, volume 22 of 12, pages 1349-1380. IEEE, December 2000.

[31] RQL: A Functional Query Language for RDF, G. Karvounarakis, A. Magkanaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, K. Tolle, to be published at Functional Approaches to Computing With Data, P.M.D.Gray, L.Kerschberg, P.J.H.King, A.Poulovassilis (eds.), LNCS Series

[32] RDF Resource Description Framework http://www.w3c.org/RDF/

[33] The Harmony Project, http://metadata.net/harmony

[34] M doer, J Hunter, C Lagoze "Towards a Core Ontology for Information Integration" Journal of Digital Information volume 4, issue 1. April 2003

[35] IconClass http://www.iconclass.nl/

[36] Art & Architecture Thesaurus from the Getty Trust http://www.getty.edu/research/tools/vocabulary/aat/

[37]CIDOC CRM expressed in DAML http://thesauri.dbalzer.net/cidoc-crm/daml/

[38]CIDOC CRM and MPEG 7 http://archive.dstc.edu.au/RDU/staff/jane-hunter/MW2002/paper.html

[39] Ontoweb http://ontoweb.aifb.uni-karlsruhe.de/