# 2002-08-12

# D6.2 Impact on World-Wide Metadata Standards

| | | | | |
|---|---|---|---|---|
| **Project acronym** | ARTISTE | | | |
| **Contract number** | IST 11.978 | | | |
| **Deliverable number** | D6.2 | | | |
| **Deliverable title** | D6.2 Impact on World-Wide Metadata Standards | | | |
| **Workpackage** | WP6 Distributed Query and Metadata Standards | | | |
| **Task** | T6.4 Metadata Impact on Standards | | | |
| **Date of delivery** | **Contractual** | PM24 | **Actual** | 2002-08-12 |
| **Code name** | 905-0004655 Rev. A | | **Version 2.0** | draft ☐ final ☒ |
| **Nature** | Report | | | |
| **Dissemination level** | Public | | | |
| **Authors (Partner)** | IT Innovation Centre | | | |
| **Contact Person** | Matthew Addis<br>Alison Stevenson<br><br>IT Innovation Centre<br>2 Venture Road<br>Chilworth Science Park, Southampton<br>S016 7NP, United Kingdom | | mailto:mja@it-innovation.soton.ac.uk<br>mailto:as@it-innovation.soton.ac.uk<br><br>Tel: +44 23 8076 0834<br>Fax: +44 23 8076 0833<br><br>http://www.it-innovation.soton.ac.uk | |
| **Abstract** | This document presents the ARTISTE three-level approach to providing an open and flexible solution for combined metadata and image content-based search and retrieval across multiple, distributed image collections.<br><br>The intended audience for this report includes museum and gallery owners who are interested in providing or extending services for remote access, developers of collection management and image search and retrieval systems, and standards bodies in both the fine art and digital library domains. | | | |

## Document Changes

| Rev. | Date | Section | Comment |
|---|---|---|---|
| A | 2002-08-12 | All | Initial Issue |

## Reviewers of Current Revision

| Rev. | Name, Organization | Role |
|---|---|---|
| A | ARTISTE PMT | Project Management Team |
| | Warren Sterling | Project Board Chair |

## Conventions Used in This Document

The following notational conventions are used in this document:

- Variable and Styles names are shown in Arial 9 pt: Variable1
- Code is shown in Courier New 10 pt: `While True Do`
- Commands are shown in Courier New 11 pt: `Delete`

## Trademarks



## Control Information

Page 43 is the last page of this document.

# Executive Summary

ARTISTE is a European Commission supported project that has developed integrated content and metadata-based image retrieval across several major art galleries in Europe. Collaborating galleries include the Louvre in Paris, the Victoria and Albert Museum in London, the Uffizi Gallery in Florence and the National Gallery in London.

Museums and galleries often have several digital collections ranging from public access images to specialised scientific images used for conservation purposes. Direct access from one gallery to another is currently uncommon for textual data and almost unheard of in terms of image-based search and retrieval. However, cross-collection access is recognised as important, for example to compare the treatments and conditions of Europe's paintings, which form a core part of our cultural heritage.

Over the last two and a half years, ARTISTE has developed an image search and retrieval system that integrates distributed, heterogeneous image collections. This report positions the work achieved in ARTISTE with respect to metadata standards and approaches for open search and retrieval using digital library technology.

In particular, this report describes three key aspects of ARTISTE:

- The transparent translation of local metadata schema to common standards such as Dublin Core and CIMI consortium attribute sets to allow cross-collection searching;
- A methodology for combining metadata and image content-based analysis into single search queries to enable versatile retrieval and navigation facilities within and between gallery collections; and
- An open interface for cross-collection search and retrieval that advances existing open standards for remote access to digital libraries, such as OAI (Open Archive Initiative) and ZING SRW (Z39.50 International: Next Generation Search and Retrieval Web Service).

A large part of ARTISTE is concerned with use of existing standards for metadata frameworks. However, one area where existing standards have not been sufficient is multimedia content-based search and retrieval. A proposal has been made to ZING for additions to SRW. This will hopefully enable ARTISTE to make a valued contribution to this rapidly evolving standard.

The intended audience for this report includes museum and gallery owners who are interested in providing or extending services for remote access, developers of collection management and image search and retrieval systems, and standards bodies in both the fine art and digital library domains.

# Contents

# 1. Introduction

## 1.1 ARTISTE Project Overview

The ARTISTE project [1], partly funded by the EU under the fifth R&D framework, has developed a system for the automatic indexing and cross-collection search and retrieval of high-resolution art images.

Four major European galleries are involved in the project: the Uffizi in Florence, the National Gallery and the Victoria and Albert Museum in London, and the Centre de Recherche et de Restauration des Musées de France (C2RMF) which is the Louvre related restoration centre. The ARTISTE system currently holds over 160,000 images from four separate collections owned by these partners.

The galleries have partnered with NCR, a leading player in database and Data Warehouse technology; Interactive Labs, the new media design and development facility of Italy's leading art publishing group, Giunti; IT Innovation, a specialist in building innovative IT systems; and the Department of Electronics and Computer Science at the University of Southampton.

After two and a half years and over twenty person-years of effort by the partners, the ARTISTE project is now reaching a conclusion. A summary of the project describing the results achieved has recently been published in Cultivate Interactive [2].

The work started in ARTISTE is being continued in SCULPTEUR [30], another project funded by the European Commission. SCULPTEUR will develop both the technology and the expertise to create, manage and present cultural archives of 3D models and associated multimedia objects.

## 1.2 Rationale

European museums and galleries are rich in cultural treasures but public access to internal digital image collections for education, leisure or work purposes has not reached its full potential. Automation of the indexing, search, retrieval and delivery of such assets over the web would help improve accessibility, in turn broadening public awareness of the European cultural heritage that lies behind them.

Realising the potential of digital image collections of cultural heritage forms a core part of the objectives of the IST III.2.3 ‘Access to scientific and cultural heritage’ action line under which ARTISTE is funded.

*To improve access by citizens and by professionals to Europe’s fast-growing science and cultural knowledge base through developing advanced systems and services supporting*

*large scale distributed, multi-disciplinary collections of cultural and scientific multimedia resources. (Component of IST Action line III.2.3 objective)*

*Work should also address interoperable access to distributed resources, whether through cross-domain resource discovery, interfaces or new architectures and standards, or whether through digital archives integrating library and museum objects. (Component of IST Action line III.2.3 objective)*

The above objectives are directly reflected in the wide range of image indexing, search and retrieval requirements of the ARTISTE end users. These requirements include:

- The management and documentation of image collections by museums curators including cross-referencing to collections in other museums and galleries or external information sources;
- Cross-collection search and retrieval used for research, decision taking and co-operation amongst different teams of researchers, historians and conservationists;
- Quick access and retrieval of images and related information for documentation or inclusion in the products of publishing and editing organisations;
- Retrieval of precise information on the art market such as history of ownership, location and transfers by auctioneers, art experts, and art amateurs; and
- Products and services built on remote access to image collections for education and public end uses.

Remote and cross-collection access to digital image information is a key objective of ARTISTE. This includes achieving interoperability between ARTISTE accessible collections and other digital library resources. Since image collections are typically stored in separate databases and all have their own unique schema for the meta-data that describes their contents, ARTISTE also needs to provide quick and transparent searching of multiple collections as if they were a single entity.

## 1.3 Approach

The exploitation of cultural image collections is limited because of a lack of relevant metadata to describe the images, lack of conformance to common schema for metadata that does exist, and lack of appropriate and convenient access methods. The ARTISTE project seeks to address these issues.

The approach we have taken in ARTISTE is to use existing standards and technologies where possible for metadata structuring and translation. This in turn underpins an open standards approach to providing an open interface for metadata harvesting and image search and retrieval.

ARTISTE makes considerable use of existing open metadata standards such as Dublin Core [3] and RDF Schema [4]. Access to this metadata is supported through the Open Archive Initiative (OAI) [5] information retrieval standard for distributed access. Open query and retrieval of images across multiple collections is then provided by a Search and Retrieve Web service (SRW) [7] based on z39.50 [6].

The goal of the OAI harvesting protocol is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the Web. ARTISTE is an OAI data provider and has

implemented support for the Open Archives Initiative Protocol for Metadata Harvesting, thus providing open access to metadata stored with each museum and gallery collection.

ARTISTE is also participating in an initiative to redesign the primary open standard for interoperability between digital libraries, z39.50, using web technologies such as XML and SOAP. The z39.50 into the Next Generation (ZING) [8] initiative has proposed a Search and Retrieve Web Service based on the z39.50 protocol for searching databases that contain metadata and objects.

ARTISTE is one of the early implementers [9] of SRW and has extended the capabilities of SRW to enable image content and metadata based searches over multiple ARTISTE collections. Having emerged from the digital library community, z39.50 has been traditionally concerned with text based searching. ARTISTE has been working with ZING to incorporate the ability to deal with content-based searching of images and thus expand international standards of information retrieval.

Throughout the project, ARTISTE has tracked appropriate standards and has provided feedback to the developers and users of those standards on how they could be improved to support images as well as textual information.

## 1.4 Report Structure

The remainder of this report is divided into the following sections.

**Section 2** presents an overview of the ARTISTE architecture for cross-collection search and retrieval. This includes a discussion of how image collections are integrated into ARTISTE.

**Section 3** discusses the use of RDF technology to support mapping between local and common metadata schemas. This mapping allows interoperability between galleries at the metadata level without the need to change local metadata organisation and semantics.

**Section 4** describes a method for combining searches over textual metadata with searches over image content into a single query format. This vastly simplifies specification of complex searches whilst at the same time giving the underlying systems the flexibility to choose the best search techniques to use.

**Section** 5 presents the standards-based search and retrieval services provided by ARTISTE. Particular attention is paid to how ARTISTE has identified deficiencies in current standards, devised solutions and then fed these solutions back into the standards community.

**Section** 6 provides observations on existing standards and discusses some of wider considerations when building an open search and retrieval system that supports images and content-based analysis.

**Section** 7 describes how ARTISTE has contributed to current standards

**Section** 8 summarises the findings of this report into a set of conclusions

# 2. The ARTISTE architecture for cross-collection search and retrieval

## 2.1 Architecture Overview

ARTISTE uses a multi-tier architecture as shown in Figure 1.

The architecture consists of software layers separated by a series of APIs (Application Programming Interfaces) that enable loose coupling to be achieved between these main architectural components.

The use of documented and published APIs is important to allow components to be developed by different organisations.

The use of open standards such as OAI and SRW for certain interfaces provides the basis of interoperability with other software systems, for example digital libraries, that haven't been specifically developed to work with ARTISTE.

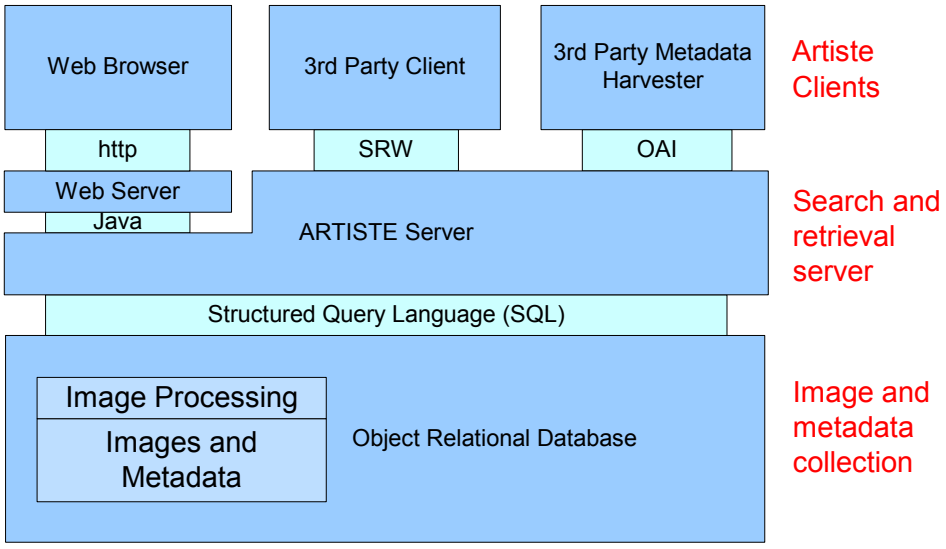


*Figure 1 ARTISTE logical architecture*

## 2.2 Image and metadata collections

Images of the art objects in a museum or gallery collection are held in an Object Relational Database Management System (ORDBMS) from NCR called Teradata Object Relational (TOR). Images are stored as Binary Large Objects (BLOBS).

Textual metadata is also stored in the database alongside the images. The metadata is stored using whatever relational schema already exists for a gallery or museum. Storing both the images and metadata in a database allows advantage to be taken of the scalability and robustness of database management systems. This is important in light of the size of the image collections involved in the project. For example, one of the collections held in ARTISTE has over 100,000 images, each of which has over 50 different metadata attributes.

ARTISTE uses a wide variety of image processing algorithms [2]**[10][11][12]** as the basis of content-based retrieval. Each algorithm is applied to the images in the collection to generate a set of image content descriptors called 'feature vectors'. A feature vector can be considered as a way of indexing an image to describe an aspect such as colour distribution or texture. The feature vectors are then integrated and stored with the text metadata for each image in the database.

When a content-based search needs to be made, the required algorithm is run on the query image to create a query feature vector. For example, the user might have a query image of a particular object that they wish to locate in a collection. Alternatively, the user might have a query image containing a particular range of colours, e.g. a certain pigment, where they are interesting finding images with similar colours in the image collection.

The query feature vector is then compared with all the corresponding feature vectors for the images in the collection. The comparison of feature vectors results in a measure of distance between the query image and each image in the collection. The images in the collection are then returned to the user as a series of thumbnails in order of increasing distance. In some cases, the algorithms can be combined into composite queries and a normalised distance measure for each algorithm is used to determine the overall match of a result image to the query. The image processing algorithms are executed in the database by wrapping them as User Defined Modules (UDM) that can be called directly from SQL queries.

Creating all the feature vectors in advance for a collection of images greatly improves search time since they do not need to be created every time a content-based search is performed.

## 2.3 ARTISTE Server

The ARTISTE server uses SQL queries to access the textual metadata, feature vectors, images and image processing algorithms in the object relational database. The server is responsible for translating queries from the client applications into SQL that is appropriate for the local metadata and image schema. The ARTISTE server is also able to communicate with other ARTISTE servers so that the query can be executed across multiple, distributed collections.

The server is implemented as an Enterprise Java application.

## 2.4 ARTISTE Clients

ARTISTE supports access by clients using several different protocols:

- Web based access can be made from a standard browser as ARTISTE supports a simple, wizard driven user interface that makes it easy to create, execute and browse the results of complex metadata and content-based queries.

- Metadata can be extracted from the ARTISTE server through support for the OAI Metadata Harvesting protocol. This only allows the textual metadata on the images to be retrieved.
- Third party software applications can execute metadata and content-based searches through the SRW interface. These applications could be third-party browsers, eLearning applications, other digital library systems, and other ARTISTE servers.

## 2.5 Deployment

The use of a multi-tier architecture allows physical distribution of the system components so that cross-collection searching can be performed when the image collections are physically located at each of the gallery sites. This is achieved by having a local image database and ARTISTE server at each site that owns an image collection.

The distributed architecture gives the galleries control over their own collections and the support for legacy database schemas keeps to a minimum the effort required by galleries make those collections open to retrieval, navigation and interoperability with other collections.

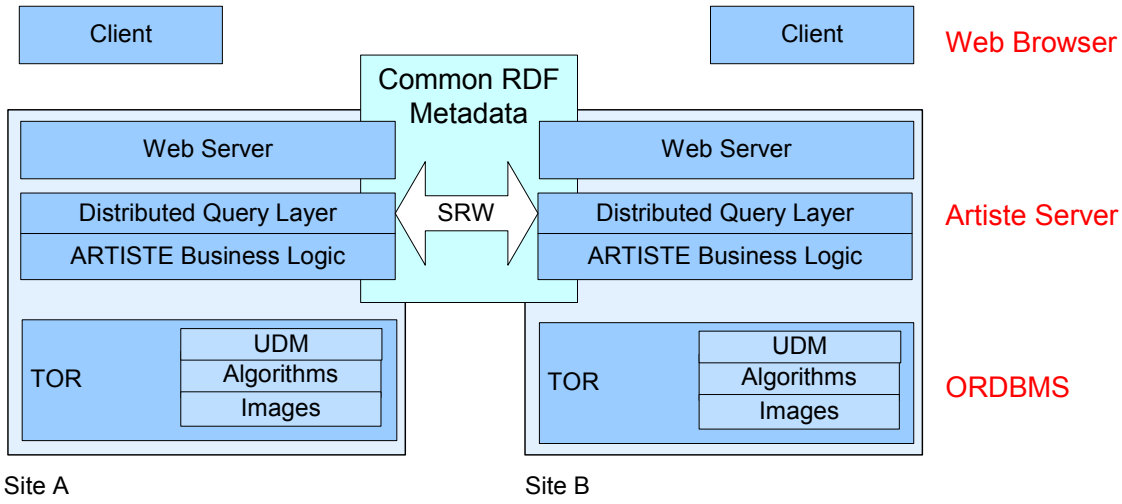


*Figure 2 ARTISTE physical architecture*

The ARTISTE servers communicate with each other through a 'distributed query layer' that uses the SRW (Search and Retrieve Web service) protocol as shown in Figure 2. The distributed query and metadata layer provides a single interface to the art, its metadata, and facilities to enable queries to be directed towards multiple distributed databases. SRW provides a framework for search and retrieval but isn't in itself sufficient to provide interoperability. A common metadata schema is required to define the semantics of the queries made between servers and the data that is returned in response. This common schema is structured using RDF and is mapped to the local metadata schemas of the individual collections. This mapping is described in more detail in Section 3

# 3. Translation to a common metadata schema

## 3.1 Introduction

In ARTISTE, there is no requirement for the pre-existing metadata for image collections to conform to a single schema. The metadata is a combination of pre-existing data loaded from gallery legacy systems and metadata generated directly by the ARTISTE system.

Whereas previous European research projects on Art such as AQUARELLE [33] used a standard metadata format to integrate collections, the ARTISTE system maintains the individual database schemas of the galleries.

ARTISTE enables metadata queries to be executed across the multiple, diverse collections by using Resource Description Format (RDF) to define the syntax and semantics for standard metadata terms.

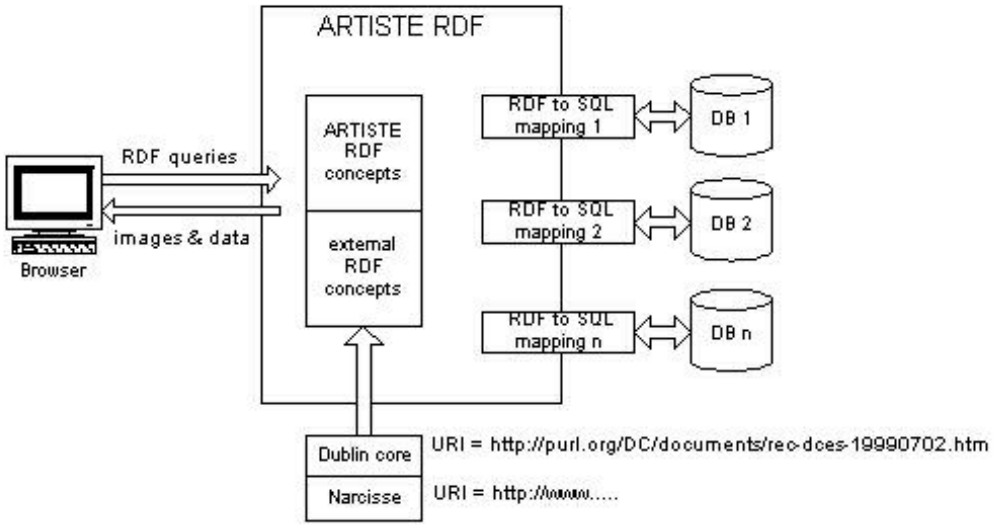


*Figure 3Use of RDF in the ARTISTE system*

The diagram in Figure 3 shows how a client issues a query to ARTISTE that is expressed in RDF, which is then processed in the context of both internal and external concepts, and mapped to a set of SQL queries to be executed against the databases containing the image collections and metadata. The resulting images and metadata that match the query are then aggregated across the collections and returned to the user as a single result set.

## 3.2 Mapping metadata using RDF

Each of the different database schemas employed by the galleries is encoded as an RDF schema, using RDF syntax [13] and the basic semantic expressions of the RDF Vocabulary Description Language 1.0 [14] more commonly known as RDFS.

The schemas, which have been validated according to the RDF Model and Syntax Specification using the W3C RDF Validation Service [15], are available expressed as XML [16]

One of the common metadata schemas supported by ARTISTE is The Dublin Core Metadata Initiative (DCMI) [3] who have defined a metadata element set consisting of 15 elements [17], for example, 'Title', 'Creator', 'Subject' and 'Date'.

In the example shown in Figure 4, shows how interoperability between two collections can be achieved by mapping the Dublin Core terms. In this case, Dublin Core 'Title' is mapped to two different columns in the collection metadata tables. Collection A defines image title within the 'Caption' column and Collection B defines image title within the 'TitleM' column.

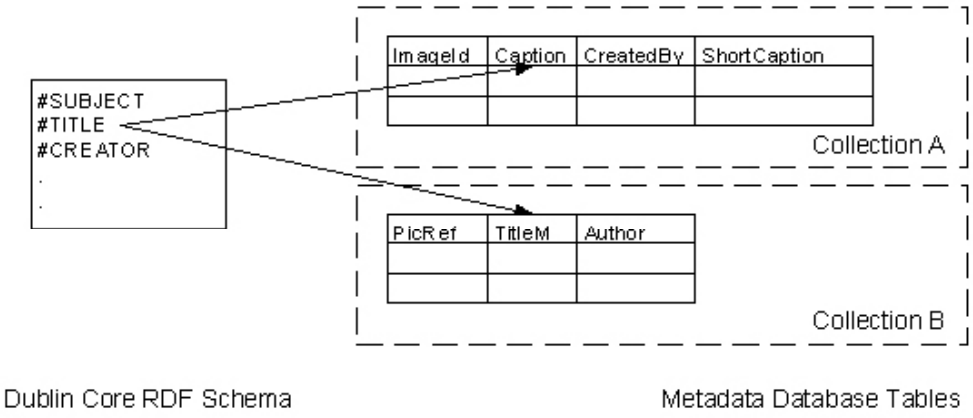


*Figure 4 Application of Dublin Core in ARTISTE*

An extract from the RDF for Collection A (In this case, the Victoria and Albert Museum Collection) is shown in Figure 5.

```
<ac:TextAttribute rdf:ID="Caption">
  <rdfs:isDefinedBy rdf:resource="http://artiste.it-
     innovation.soton.ac.uk/rdf/vam/vam.rdf.rdf#Caption"/>
  <rdfs:isDefinedBy rdf:resource=" http://artiste.it-
     innovation.soton.ac.uk/rdf/dc.rdf.rdf##title"/>
</ac:TextAttribute>
```

*Figure 5 RDF mapping of Dublin Core 'Title' for the Victoria and Albert Museum*

It is the responsibility of each site to define which of the common metadata terms it supports. In general, it is unlikely that collections owners will support the same metadata and hence querying capabilities. For example, each collection might only supports a subset of the Dublin Core and/or some additional terms defined in another standard metadata schemas.

To ensure interoperability, the ARTISTE system generates a 'Query Context' when a user builds a new query based upon the collections they wish to search. The Query Context is constructed from the RDF and contains an intersection of query capabilities supported by the selected collections. This means that the user is constrained to only query within the common capability of the collections they wish to search.

## 3.3 Multilingual metadata

Interoperability between collections in a European context is not a purely technical problem. Any system seeking to provide seamless access to multiple geographically and culturally distributed collections must take account of the multilingual nature both of the metadata and the requirements of users who wishes to access that metadata.

The ARTISTE system contains metadata in English (from the Victoria & Albert Museum and the National Gallery), French (from the C2RMF) and Italian (from the Uffizi). Furthermore the web interface to ARTISTE supports four languages (English, French, Italian and Danish) reflecting the language needs of the members of the consortium.

ARTISTE includes multiple labels and comments for each query item defined in the Collection schemas and uses the XML special attribute 'xml:lang' to specify the language used in the content of the element. The values of the attribute are language identifiers as defined by the IETF Standard RFC 1766, 'Tags for the Identification of Languages' [18].

One example of multilingual tagging is shown in Figure 6, which shows how the label 'scale' has been translated into four different languages.

```
<ac:TextAttribute rdf:ID="scale">
    <rdfs:label xml:lang="en">en:Photograph
      scale</rdfs:label>
    <rdfs:label xml:lang="fr">fr:Echelle</rdfs:label>
    <rdfs:label xml:lang="it">it:Escala</rdfs:label>
    <rdfs:label xml:lang="dk">dk:Skala</rdfs:label>
    <rdfs:comment xml:lang="en">en:The scale of the image
      with regard to the work of art eg 1.5 to 1
    </rdfs:comment>
    <rdfs:comment xml:lang="fr">fr:Echelle de l'image par
      rapport a la dimension de l'oeuvre d'art ex : 1.5 a 1
    </rdfs:comment>
    <rdfs:comment xml:lang="it">it:Scala dell'immagine in
      rapporto all'opera d'arte, per esempio 1.5 a  1
    </rdfs:comment>
    <rdfs:comment xml:lang="dk">dk:Skalen af billedet med
      henblik pa arbejdet f.eks 1.5 a 1
    </rdfs:comment>
    <rdfs:isDefinedBy rdf:resource=http://artiste.it-
      innovation.soton.ac.uk/test_rdf/c2rmf/c2rmf.rdf#scale
      />
  </ac:TextAttribute>
```

*Figure 6 Use of RDF to provide multilingual descriptions*

The example shows how ARTISTE deals with multilingual descriptions of the different metadata attributes (Title, Author, Scale etc.). Since the number of attributes for a collection is relatively small, it is at least feasible to imagine translating each attribute name into multiple languages without too much hardship. Indeed, this approach is easily

supported for the common metadata schemas and doesn't require individual galleries to make any changes or additions to their metadata.

However, this approach isn't viable for the values of each metadata attribute. For example, it would be unreasonable to expect every gallery to translate the titles of all its paintings into multiple languages. Therefore, there are issues in supporting multilingual query against metadata content.

For example, suppose a user wants to search across the National Gallery, the Uffizi and the C2RMF collections for images called "The Forest". The RDF mapping between the legacy database schemas enables the user to execute a single query across all three sites, by searching the those fields mapped to the Dublin Core term Title.

However since the metadata stored in the C2RMF database is written in French and the metadata in the Uffizi database is written in Italian, there will be no records in those databases containing the text string 'Forest' (instead those databases would contain 'Forêt' and 'Foresta' respectively).

One approach would be to translate all the metadata from the various galleries into a common language. However, this would negate one of the key advantages of the ARTISTE system, i.e. it does not require changes to be made to legacy metadata.

A possible solution is to maintain the legacy metadata in its original languages and perform translation on a query string supplied by the user, according to which database the query is being executed over. For example, if a user wanted to search multiple collections for images with title containing the word 'Forest', then ARTISTE would build and execute a query for title contains 'Forest' against the Victoria & Albert Museum, 'Foresta' against the Uffizi, and 'Foret' against the C2RMF collections. Freely accessible Web Services already exist to perform free text translation, for example using AltaVista's BabelFish available through xMethods [19]

## 3.4 Thesauri

In addition to mapping legacy database schemas to common metadata standards, and enabling multilingual functionality, ARTISTE uses RDF and RDFS to support metadata thesauri. Thesauri are controlled vocabularies that limit the number of allowed values for a particular metadata attribute, for example 'Author' can only be one of 'Monet', 'Manet' or 'Renoir'.

ARTISTE supports a total of 15 thesauri for legacy gallery metadata attributes by building on the CERED/NBII draft RDF Server Standard Documentation [20]: seven from the Victoria & Albert Museum and eight from the C2RMF.

For simple controlled vocabularies, the list of allowed terms for a given attribute is directly encoded in RDF using the resources defined in the CERED/NBII schema. This RDF document, or Collection Thesaurus is included by reference in the Collection schema for the gallery collection.

In some cases, codes are used for the values in a constrained metadata attribute. For Controlled Vocabularies using Codes, the Collection Thesaurus document contains a list of allowed codes for a given attribute, encoded in RDF using the resources defined in the CERED/NBII schema. The values these codes refer to are stored in a separate RDF document, again encoded using the resources defined in the CERED/NBII schema. This

single attribute thesaurus is included by reference in the Collection Thesaurus, which is in turn included by reference in the Collection schema for the gallery collection.

Multilingual Controlled Vocabularies are simply a special case of Controlled Vocabularies using Codes. Multilingual vocabularies consist of a list of allowed attribute values (this may be a code or the actual value in a default language representation) encoded in the collection thesaurus document. The values referred to are stored in a separate RDF document utilizing the RDFS attribute 'rdfs:label' and the XML attribute 'xml:lang' as well as the using the resources defined in the CERED/NBII schema.

All eight of the thesauri for the C2RMF metadata are multilingual, supporting English and French representations.

Figure 7 shows the multilingual values for the codes in the C2RMF metadata attribute 'Category'.

```
<?xml version="1.0" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#" xmlns:rdfs="http://www.w3.org/TR/rdf-
    schema#" xmlns:ac="http://artiste.it-
    innovation.soton.ac.uk/rdf/ArtisteCore.rdf#"
    xmlns:aconf="http://artiste.it-
    innovation.soton.ac.uk/rdf/ArtisteConfiguration.rdf
    #" xmlns:z19="http://artiste.it-
    innovation.soton.ac.uk/rdf/thesaurus.rdf#">
  <z19:Category rdf:ID="DE">
      <rdfs:label
        xml:lang="en">en:drawing</rdfs:label>
      <rdfs:label
        xml:lang="fr">fr:dessin</rdfs:label>
    </z19:Category>
  <z19:Category rdf:ID="DP">
      <rdfs:label xml:lang="en">en:drawing and
        painting</rdfs:label>
      <rdfs:label xml:lang="fr">fr:dessin et
        peinture</rdfs:label>
    </z19:Category>
  <z19:Category rdf:ID="EN">
      <rdfs:label
        xml:lang="en">en:illumination</rdfs:label>
      <rdfs:label
        xml:lang="fr">fr:enluminure</rdfs:label>
    </z19:Category>
```

*Figure 7 Different codes for 'Category' and their values in French and English*

The controlled vocabulary that constrains the number of codes for 'Category' is shown in *Figure 8*

```
<?xml version="1.0" ?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#" xmlns:rdfs="http://www.w3.org/TR/rdf-
    schema#" xmlns:ac="http://artiste.it-
    innovation.soton.ac.uk/rdf/ArtisteCore.rdf#"
    xmlns:aconf="http://artiste.it-
    innovation.soton.ac.uk/rdf/ArtisteConfiguration.rdf#
    " xmlns:z19="http://artiste.it-
    innovation.soton.ac.uk/rdf/thesaurus.rdf#"
    xmlns:categ="http://artiste.it-
    innovation.soton.ac.uk/rdf/c2rmf/c2rmf-
    thesaurus/categ.rdf#">
  <z19:Category rdf:ID="categ">
      <z19:IC>DE</z19:IC>
      <z19:IC>DP</z19:IC>
      <z19:IC>EN</z19:IC>
      <z19:IC>ES</z19:IC>
      <z19:IC>OR</z19:IC>
      <z19:IC>PE</z19:IC>
      <z19:IC>SC</z19:IC>
      <z19:IC>ST</z19:IC>
      <z19:IC>TA</z19:IC>
    </z19:Category>
```

*Figure 8 Controlled Vocabulary using codes for the metadata attribute 'Category'*

## 3.5 Summary

The ARTISTE system enables queries to be executed across multiple, distributed collections without requiring each collection to conform to a standard schema.

RDF is used to define the syntax and semantics for standard metadata terms. Each collection provides a mapping that relates these standard metadata terms to individual database table and column values. Where appropriate, IT Innovation has contributed RDF versions of established schema back to the community.

Queries are composed using RDF, and subsequently translated to SQL at each site. Benefits of this approach include:

- The use of RDF mapping provides a flexible solution to cross-collection searching;
- Mapping to a common schema allows common semantics to be supported without needing changes to local metadata and schemas;
- Users can be dynamically constrained in their querying through use of a 'Query Context' so that they only request queries that are within the common capabilities of a set of collections.
- Multilingual translation of metadata attribute names allows the user to use their native language when specifying which attributes to search over for multiple collections.
- Free text translation is a possible solution to multilingual searching of metadata content.

Achieving interoperability between gallery collections through RDF is not itself enough to guarantee improved access to digital image resources. A standard way of specifying queries and then performing search and retrieval according to well-understood protocols is still required. These aspects are covered in the next two sections of this document

# 4. Integrated metadata and image content-based analysis

## 4.1 Introduction

A key feature of ARTISTE is the ability to provide image content-based querying in addition to searches based on textual metadata.

Current digital library query representations and protocols, such as Z39.50, deal entirely with textual metadata. This is not sufficient for multimedia digital libraries such as ARTISTE, where searches can be made on image content as well as textual metadata. In particular, current protocols have the following restrictions.

1. There are no methods for specifying image content as a metadata item.
2. There are no operators defined relating to image content.
3. There are no methods for carrying out searches that result in the execution of image processing algorithms.

ARTISTE has addressed each of these issues by using RDF and RDFS to define a query ontology which defines and describes the objects, methods and operators required to build a query based on either image content, textual metadata or a combination of the two.

## 4.2 Specifying Image Content

All items that can be queried in ARTISTE are described by a 'Query Item' hierarchy, as shown in Figure 9 A Query Item can be an image, properties of an image (such as colour or shape), or attributes associated with an image (conventional metadata such as textual and numeric items).

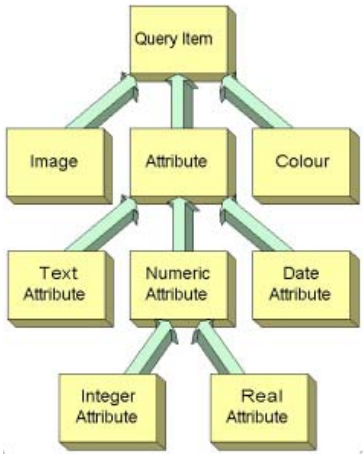


*Figure 9 ARTISTE Query Item hierarchy*

This structure is defined in the ARTISTE CORE RDF schema as shown in Figure 10.

```
<rdfs:Class rdf:ID="QueryItem">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/1999/02/22-
rdf-syntax#Property" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="Image">
  <rdfs:subClassOf rdf:resource="#QueryItem" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="Colour">
  <rdfs:subClassOf rdf:resource="#QueryItem" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="Attribute">
  <rdfs:subClassOf rdf:resource="#QueryItem" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="TextAttribute">
  <rdfs:subClassOf rdf:resource="#Attribute" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="IntegerAttribute">
  <rdfs:subClassOf rdf:resource="#Attribute" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="RealAttribute">
  <rdfs:subClassOf rdf:resource="#Attribute" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="DateAttribute">
  <rdfs:subClassOf rdf:resource="#Attribute" />
  </rdfs:Class>
- <rdfs:Class rdf:ID="FeatureVectorAttribute">
  <rdfs:subClassOf rdf:resource="#Attribute" />
  </rdfs:Class>
```

*Figure 10 ARTISTE Query Item RDF schema*

## 4.3 Query Operators

The operations that can be performed on ARTISTE Query Items are also explicitly defined in ARTISTE as Query Operators, as shown in Figure 11. Current query

operators include exact operators (such as equals, less than etc.) and fuzzy operators (such as similar to).

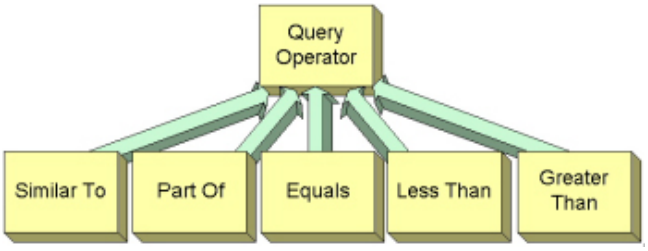


*Figure 11 ARTISTE Query Operator hierarchy*

An extract of the RDF that is used to describe the Query Operator hierarchy is shown in Figure 12.

```
<rdfs:Class rdf:ID="QueryOperator">
   <rdfs:label xml:lang="en">QueryOperator</rdfs:label>
   <rdfs:comment>An operator that can be specified in an ARTISTE
   query.
   </rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="PartOf">
   <rdfs:subClassOf rdf:resource="#QueryOperator"/>
   <rdfs:label xml:lang="en">PartOf</rdfs:label>
   <rdfs:comment>The concept of being part of another object.
   </rdfs:comment>
</rdfs:Class>
```

*Figure 12 Extract from ARTISTE Query Operator RDF schema*

## 4.4 Query Expression Rules

As described above, Query Items and Query Operators are defined in separate RDF schemas. However, not all Query Operators can be applied to all Query Items. For example, the concept of 'Greater Than' applied to 'Colour' is nonsensical.

Therefore, in ARTISTE, all operators must be instantiated to form rules – each rule describes a legal ARTISTE query expression and specifies which query items can be used with each query operator. Each query operator is assumed to take two operands: a subject and an object. Properties are also defined which govern the particular query items that a particular operator can take.

Rules governing non-fuzzy operators such rules such as DateEquals or TextContains, can be defined simply in terms of operator, object and subject.

However, in the case of image content-based queries where the operator is 'fuzzy', for example 'Part Of', the query expression is also linked to the appropriate image analysis algorithm as shown in Figure 13.

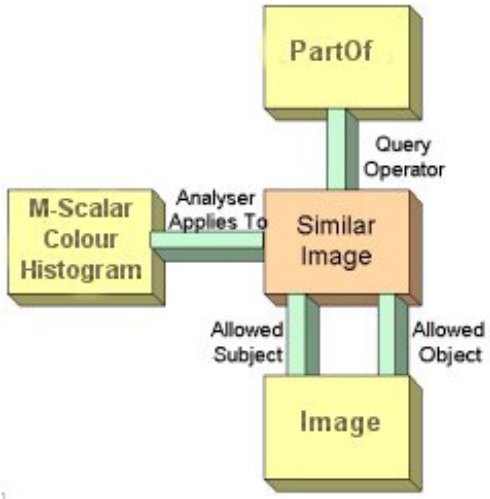


*Figure 13 ARTISTE Query Expressions*

The diagram in Figure 13 shows a diagrammatic representation of the 'SimilarSubImage' query expression rule. This rule defines 'SimilarSubImage' as an operation involving two Images and the 'PartOf' operator. The RDF that describes the 'SimilarSubImage' rule and its association with the MultiScalarColourHistogram image processing algorithm is shown in Figure 14

```
<ac:QueryExpressionRule rdf:ID="SimilarSubImage">
   <ac:QueryOperator>#PartOf</ac:QueryOperator>
   <ac:AllowedSubject>#Image</ac:AllowedSubject>
   <ac:AllowedObject>#Image</ac:AllowedObject>
   <rdfs:label xml:lang="en">Similar SubImage</rdfs:label>
   <rdfs:comment>The concept of an image being similar to part of
   another image.</rdfs:comment>
</ac:QueryExpressionRule>

<ac:Analyser rdf:ID="MultiScalarColourHistogram">
   <rdfs:label
   xml:lang="en">en:MultiScalarColourHistogram</rdfs:label>
   <rdfs:comment>Multi Scalar Colour Content</rdfs:comment>
   <ac:AnalyserAppliesTo>http://artiste.it-
   innovation.soton.ac.uk/rdf/ArtisteCore.rdf#SimilarSubImage
   </ac:AnalyserAppliesTo>
</ac:Analyser>
```

*Figure 14 Extract from the ARTISTE Query Expression RDF schema*

## 4.5 Combining textual metadata searching with image content-based analysis

This section contains an example that illustrates combined querying using textual metadata and image content-based analysis.

The extract in Figure 15 is from an RDF application profile which describes an example site. The site declares that it supports the 'MultiScalar Colour Histogram' algorithm, an

image query item for visible light images, and the 'title' text attribute (as defined by Dublin Core).

The following pseudo code represents a combined query for images in the collection that both contain the specified query image as a sub image and have the title Francis.

```
Query

Select all where

ac:VisibleLightImage ac:SimilarSubImage
ac:MultiScalarColourHistogram, 'myQueryImage'

And

dc:Title ac:TextEquals 'Francis'
```

One important aspect of the query above is that it contains no concrete mappings to actual database fields so it can be passed without alteration to multiple gallery sites. At each site it will be translated to SQL and in particular the Dublin Core Title is translated to the appropriate table and column name for the local metadata schema.

```
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/TR/rdf-schema#" xmlns:ac=http://artiste.it-
innovation.soton.ac.uk/rdf/ArtisteCore.rdf#>

<ac:Analyser rdf:ID="MultiScalarColourHistogram">
   <rdfs:label xml:lang="en">MultiScalarColourHistogram
   </rdfs:label>
   <ac:AnalyserAppliesTo>http://artiste.itinnovation.soton.ac.uk/rdf/ArtisteCore.
   rdf#SimilarSubImage
   </ac:AnalyserAppliesTo>
   <rdfs:isDefinedBy rdf:resource="http://artiste.it-
   innovation.soton.ac.uk/rdf/ArtisteConfiguration.rdf#"/>
</ac:Analyser>

<ac:Image rdf:ID="VisibleLightImage">
   <rdfs:label xml:lang="en">Visible Light Image</rdfs:label>
   <rdfs:isDefinedBy rdf:resource="http://artiste.it-
   innovation.soton.ac.uk/rdf/ArtisteCore.rdf#VisibleLightImage"/>
</ac:Image>

<ac:TextAttribute rdf:ID="Caption">
   <rdfs:label xml:lang="en">Caption</rdfs:label>
   <rdfs:isDefinedBy rdf:resource="http://artiste.it-
   innovation.soton.ac.uk/rdf/CollectionA.rdf#Caption" />
   <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/elements/1.1/title" />
</ac:TextAttribute>

</rdf:RDF>
```

*Figure 15 Extract from an ARTISTE RDF Site Profile*

## 4.6 A graphical approach to combining textual metadata searching with image content-based analysis

As a contrast to the RDF in Section 4.5 above, the series of images below show the specification, execution and results of the same combined query

Users of the ARTISTE system can search for images using the Web Browser user interface that guides them through the query formulation process using a query wizard.

The wizard prompts the user with self-explanatory and non-technical questions. The wizard permits forward and backward movements through the process of search-generation. In this way, users of ARTISTE can quickly and simply build up sophisticated queries without needing to understand the technical details about the algorithms being used and why

The specification of the query as guided by the wizard like user interface is shown in Figure 16, Figure 17, and Figure 18.

The results of the query are shown in Figure 19. All the result images have 'Francis' in the title. The result in the bottom left hand corner is the parent art object from which the query image was derived – in this case a small wooden figure of St Francis of Assisi in the Victoria and Albert Collection.

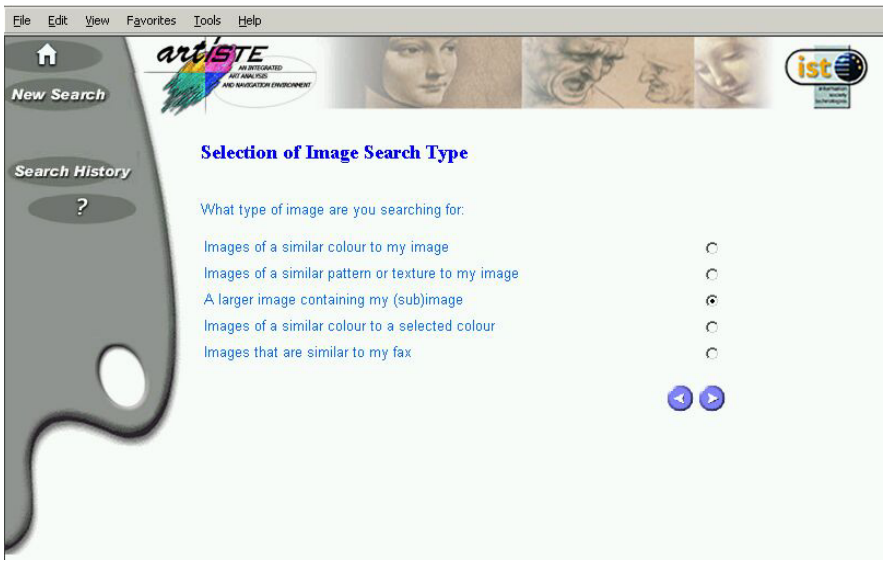


*Figure 16 Using the ARTISTE Web Interface to specify a content-based query*

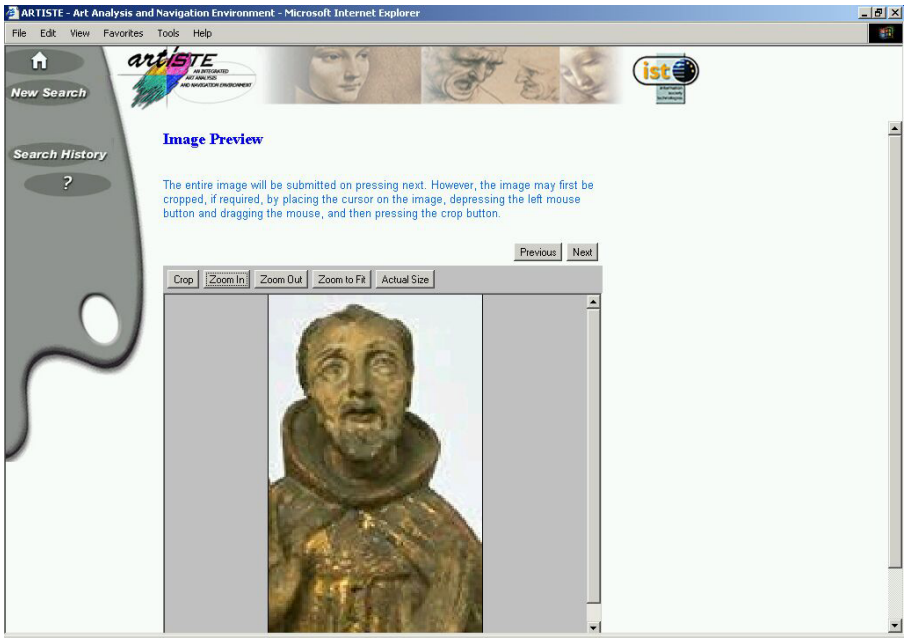


*Figure 17 Using the ARTISTE Web interface to select a query image*

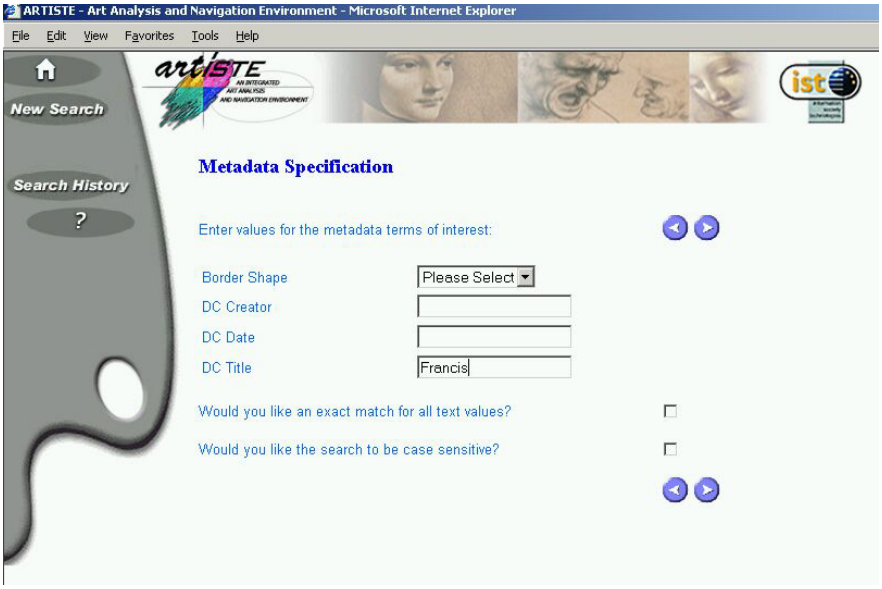


*Figure 18 Using the ARTISTE Web interface to add a textual metadata search term*

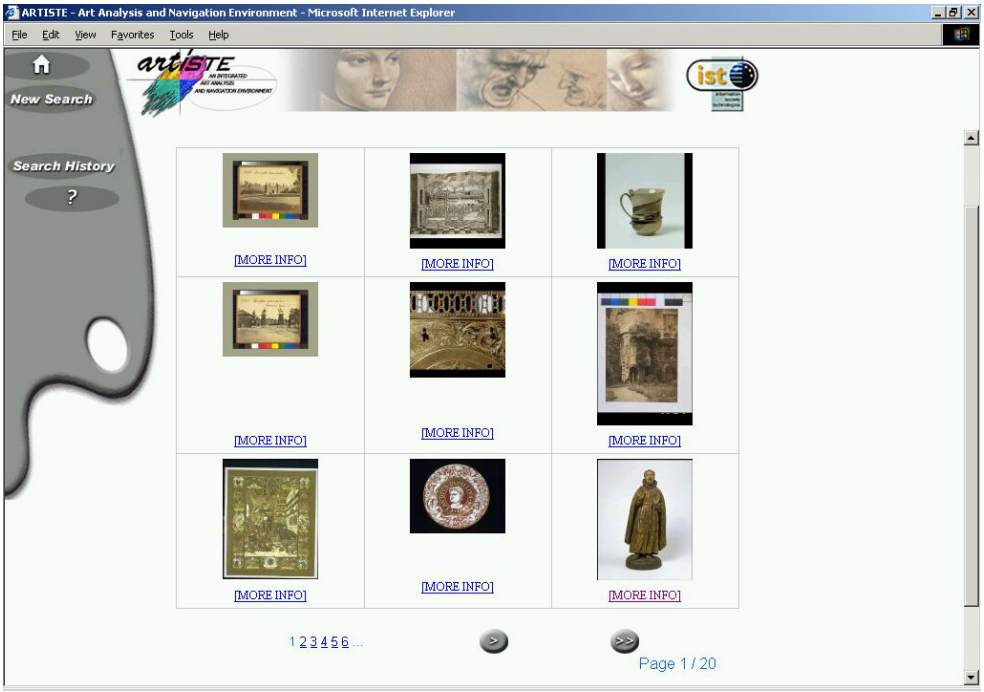


*Figure 19 Results of the combined query displayed using the ARTISTE Web interface. Note that the result in the bottom right hand corner contains the query image as a sub image.*

A publicly accessible ARTISTE demonstration system that supports Web browser based access to a 1000 image collection is hosted by IT Innovation [31]

## 4.7 Summary

A key feature of ARTISTE is the ability to provide image content-based querying in addition to searches based on textual metadata.

ARTISTE uses RDF and RDFS to explicitly:

[1] Describe image content as a metadata item.

[2] Define operators relating to image content.

[3] Capture the constraints under which operators can be applied to both image content.

Using a single framework for specifying the semantics of both image content and textual metadata based queries allows these to be declaratively and seamlessly integrated into a single query language.

However, common metadata schema as describe in Section 3 along with a framework for query items, operators and expressions is not enough to guarantee improved access to digital image resources. Well-understood protocols are still needed for executing search and retrieval. This final piece of the jigsaw is described in the next section.

# 5. Search and retrieval standards for image collections

## 5.1 Introduction

Extensive use of RDF in ARTISTE provides a way to establish common semantics between heterogeneous digital libraries containing image collections. These common semantics include how to perform content-based analysis as well as textual metadata searching. This goes a long way towards interoperability between multiple digital libraries. In fact, this is sufficient to enable cross-collection search and retrieval between ARTISTE systems. However, common semantics are not enough to provide interoperability with third-party systems. To achieve this requires adoption of standards for the process of search and retrieval itself, i.e. use of standard protocols.

ARTISTE supports two standards in this area. The first is the Open Archive Initiative (AOI) protocol for metadata harvesting. The second is a Search and Retrieval Web service standard proposed by ZING, which builds upon the well-established z39.50 digital library protocol. This section describes how ARTISTE uses and extends both of these standards.

## 5.2 Open Archives Initiative

The goal of the OAI harvesting protocol is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the Web. ARTISTE is an OAI data provider and has implemented support for the Open Archives Initiative Protocol for Metadata Harvesting, thus providing open access to metadata stored with each museum and gallery collection

In OAI terminology, each installation of the ARTISTE system acts as a metadata repository. This is defined as a network accessible server that can process the 6 OAI-PMH requests:

[1] Identify

[2] List Metadata Formats

[3] List Identifiers

[4] List Records

[5] List Sets

[6] Get Record

Thus each Site (Victoria & Albert Museum, C2RMF, National Gallery, Uffizi) is enabled to act as OAI Repository.

The OAI-PMH distinguishes between three distinct entities related to the metadata made accessible by the OAI-PMH in a repository: resource, item and record.

### 5.2.1 OAI-PMH Resource

A resource is the object that is described by the metadata. The nature of a resource, whether it is physical or digital, or whether it is stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH. In ARTISTE, the resources are the images stored in the database and aren't returned to the client as the result of an OAI-PMH request. However, it would be simple to include the URI for the images as one of the metadata attributes. The client would then be able to access the image.

### 5.2.2 OAI-PMH Item

An item is conceptually a container that stores, or dynamically generates, metadata about a single resource in multiple formats. Each format can be harvested as records via the OAI-PMH. In the ARTISTE implementation of the OAI-PMH, items are conceptually equivalent to unique identifiers that are stored in the databases for each of the images since all the metadata associated with an image can be accessed via this identifier.

Each item has an identifier that is unique within the scope of the ARTISTE OAI repository in which it is a constituent. It would be possible to use the database image_id as the item identifier for OAI. However the OAI-PMH specifies that the format of unique ID for records must correspond to that of the URI syntax. The OAI-PMH recommends that further formatting (in an XML Schema named 'oai-identifier').

To comply with this recommendation, the ARTISTE implementation includes the generation of oai-identifiers for all the images in the ARTISTE system. These take the form oai:artiste:SiteName/ImageID, for example, oai:artiste:c2rmf/13640.

### 5.2.3 OAI-PMH Record

A record is metadata in a specific format. A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item. The XML-encoding of records is organised into the following parts:

- *Header,* which contains the unique identifier of the item and properties necessary for selective harvesting. The header consists of the unique identifier of an item in a repository, and the date of creation, modification or deletion of the record for the purpose of selective harvesting.

- *Metadata,* which is a single manifestation of the metadata from an item. The OAI-PMH supports items with multiple manifestations (formats) of metadata. The ListMetadataFormats request returns the list of all metadata formats available from a repository, or for a specific item (which can be specified as an argument to the ListMetadataFormats request). ARTISTE supports the OAI Dublin Core metadata format for all of the galleries. In addition, ARTISTE can also provide metadata in the proprietary schema that is particular to each gallery.

### 5.2.4 Example metadata record

The example in Figure 20 shows an example of a metadata record returned by ARTISTE. The header contains a unique identifier of the item from which the record was disseminated (oai:artiste:Sample/46518) and the datestamp of the record (2002-02-28).

The metadata part of the record contains Dublin Core information such as 'title', 'subject' and 'description'.

```
<?xml version="1.0" encoding="UTF-8" ?>
-<GetRecord xmlns="http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_GetReco
    rd http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
    <responseDate>2002-07-31T12:37:17+01:00</responseDate>
    <requestURL>http://artiste.it-
      innovation.soton.ac.uk/servlet/OAIServlet?verb=GetRecord
    </requestURL>
  <record>
    <header>
        <identifier>oai:artiste:Sample/46518</identifier>
        <datestamp>2002-07-31T12:37:26+01:00</datestamp>
    </header>
    <metadata>
      <oai_dc xmlns="http://purl.org/dc/elements/1.0/"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://artiste.it-
          innovation.soton.ac.uk/rdf/oai_dc.rdf
          http://artiste.it-
          innovation.soton.ac.uk/rdf/oai_dc.xsd">
          <contributor>taylorc</contributor>
          <identifier>pcd228810120146-026</identifier>
          <description>An Indian black buck is shown being led
            by it's keeper. The border of the image is a floral
            design and watercolour and gold on
            paper.</description>
          <subject>INDIAN PAINTING</subject>
          <source>CT1755</source>
          <title>Indian Black BuckMughalc.1615No Date</title>
          <creator>Indian Black BuckMughalc.1615No
            Date</creator>
        </oai_dc>
      </metadata>
    </record>
  </GetRecord>
```

*Figure 20 Example ARTISTE response to an OAI request for a metadata record*

## 5.3 ARTISTE implementation and extension of the ZING Search and Retrieve Web Service

ARTISTE is participating in an initiative to redesign the primary open standard for interoperability between digital libraries, z39.50, using web technologies such as XML and SOAP. The z39.50 into the Next Generation (ZING) initiative has proposed a Search

and Retrieve Web Service (SRW) based on the z39.50 protocol for searching databases that contain metadata and objects.

ARTISTE is one of the early implementers of SRW and has extended the capabilities of SRW to enable image content and metadata based searches over multiple ARTISTE collections. Having emerged from the digital library community, z39.50 has been traditionally concerned with text based searching. ARTISTE has been working with ZING to incorporate the ability to deal with content-based searching of images and thus expand international standards of information retrieval.

### 5.3.1 Z39.50

The z39.50 protocol specifies formats and procedures governing the exchange of messages between a client and server. This enables the client to request that the server search a database and identify records that meet specified criteria, and to retrieve some or all of the identified records.

A Z39.50 server must support a core of functions: initialisation, search and retrieval. The basic process of querying is shown in Figure 21

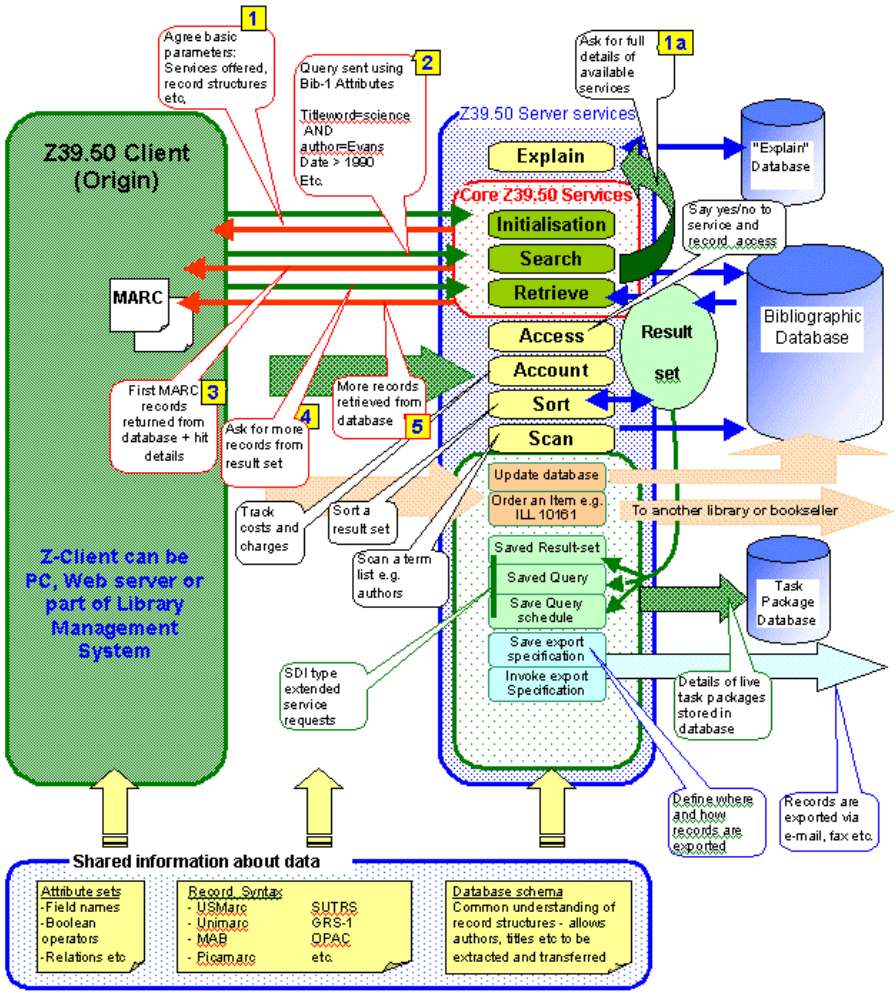


*Figure 21 Overview of z39.50 functionality. Diagram from Biblio Tech Review (http://www.biblio-tech.com/html/z39_50.html)*

### 5.3.2 Analysis of z39.50 support for distributed, content-based search and retrieval.

The search and retrieval functionality provided by the ARTISTE system can be described using four broad categories. These are:

i) Metadata Querying

ii) Image Querying

iii) Complex Querying

iv) Saving and Manipulating Queries

The z39.50 standard was reviewed against the functionality of ARTISTE with the comparison presented in Table 1

| ARTISTE functionality | Supported by z39.50 |
|---|---|
| Metadata Query | Y |
| Image Query | N |
| Query Across Multiple Databases | Y |
| Query Across Multiple Sites | N |
| Query Using Multiple Metadata Terms | Y |
| Query Using Metadata Terms and Query Images | N |
| Query Using Controlled Metadata Terms | N |
| Query Using Multilingual Metadata Terms | N |
| Save a Query | N |
| Delete a Query | N |
| Browse previous Query Runs and Retrieve Results for a Query | N |
| Support for dynamics of new & changing metadata | N |

*Table 1 Comparison of ARTISTE distributed search and retrieval functionality with z39.50*

The limited support of z39.50 for the functionality available in ARTISTE, notably the lack of support for image content-based queries, means that z39.50 is not suitable standard for exposing ARTISTE functionality to third-party software applications.

### 5.3.3 ZING Search and Retrieve Web service

ZING (Z39.50 International: Next Generation) is an umbrella term for a series of initiatives that are being pursued by the z39.50 community. The ZING initiatives aim to make the semantic content of Z39.50 more broadly available and to make Z39.50 more attractive to information providers, developers, vendors, and users, by lowering the barriers to implementation while preserving the existing intellectual contributions of Z39.50.

SRW is the ZING Search/Retrieve Web Service, which defines a draft framework for query and retrieval that takes the core search and retrieval protocol from z39.50 and specifies a Web Service implementation.

At the time of writing, the SRW specifications have draft status which means they are incomplete and under frequent revision. For example, SRW recently evolved to support fuzzy operators, which is clearly of relevance to content-based queries as described in Section 4.4. A pre-release of version 1.0 is expected to be available in mid-August 2002

and release 1.0 will be announced in early October. It will remain stable for a nine-month implementation-experience period.

### 5.3.4 SRW features that differ from z39.50

The SRW protocol retains many concepts from the traditional z39.50 protocol including Result Sets, Abstract Access points, Abstract Record schemas, Explain and Diagnostics. However it differs significantly from z39.50 in several ways which make it a better candidate for the ARTISTE DQL that 'pure' or 'traditional' z39.50:

- *Result Set Named by Server*. In SRW, the server assigns a result set id. This coincides with the ARTISTE concept of a result set which persists in the underlying database and can be accessed multiple times.

- *Connections, Sessions, State*. Each invocation of the Search/Retrieve service will be a request/response sequence, via an XML/SOAP/RPC message using HTTP POST. The use of XML and SOAP makes it much simpler to develop SRW clients and servers than z39.50 Z-Client and Z-Servers since tools that support these technologies are readily available off the shelf.

- *No distinction between server and database*. SRW does not distinguish between a server and a database. A single SRW request can therefore be sent to multiple sites and/or multiple databases.

- *Single record syntax*. All SRW records are retrieved according to a single record syntax (XML). The Z39.50 concepts of element set/specification and schema are represented by XML schemas. The use of XML as a record syntax means that the SRW responses can be encoded in RDF/RDFS and thus interoperate with the ARTISTE query ontology defined in ARTISTE Core as well as the various Collection schemas.

- *String Query*. SRW specifies string queries. The query language, CQL ("Common Query Language"), is a human-readable-string query-representation based loosely on CCL (however just the query, no commands) with access points defined.

- *Flat Access Points*. Flat access points are defined, rather than utilizing attribute vectors. Again this means a better match with the ARTISTE RDF access points which do not use attribute vectors.

- *Static Explain*. Explain information is static. The ARTISTE system was not designed with the provision of a z39.50 Explain function but the provision of a static set of data about the system is implementable within the existing architecture.

- *XML instead of ASN.1*. XML is used for abstract syntax as well as encoding. ASN.1/BER is not used. Again the use of XML maps well into the ARTISTE architecture.

The most notable of the differences between SRW and z39.50 is that queries are specified using the 'Common Query Language', and XML messages are exchanged between client and server using SOAP. The use of XML and SOAP makes it much simpler to develop SRW clients and servers since tools that support these technologies are readily available off the shelf.

### 5.3.5 Limitations of SRW

The SRW initiative is primarily concerned with lowering the barriers to implementation of z39.50, it still does not address all of deficiencies in z39.50 as a basis for an ARTISTE distributed search and retrieval system as outlined above. Therefore, IT Innovation has made extensions to SRW, notably in the Common Query Language. Some of the SRW limitations are discussed in more detail below.

SRW does not support server-managed distribution of queries across multiple sites. Therefore, it would be necessary to introduce an ARTISTE Distributed Query Layer Protocol (DQLP) that carries requests to an intermediate server that manages the breakdown of the DQLP query into multiple z39.50 SRW requests. The DQLP would most likely use SOAP as a transport protocol. The intermediate server would also be responsible for the collation of results into a results set, and the storage of queries and result sets.

It is unlikely that multi-lingual or controlled metadata terms can be incorporated in SRW, firstly because their use depends on control of the client, which is not something that is part of the ARTISTE, and secondly because the SRW specification does not include services to query the server about which query items are supported. The 'Explain' functionality in SRW is based on static information only and could not interrogate the ARTISTE Query Context to discover lists of controlled values or multilingual labels for metadata attributes. This information is available to ARTISTE users via the published RDF Collection schemas.

Having emerged from the digital library community, z39.50 has been traditionally concerned with text based searching. As a result the CQL query language proposed by the SRW specification supports metadata based querying but makes no provision for content-based queries. For example there is no provision to specify image analysers in combination with image operators (see Section 4), and the search term is always assume to be a text string. Conversely, the CQL specification also allows for the formulation of more complex text queries than ARTISTE is capable of processing. For example the specifying left and/right truncation of terms is not possible within ARTISTE nor is the specification of proximity expressed in terms of "word", "sentence", "paragraph", or "element". Such functionality is not available in the TOR database that underpins the current ARTISTE implementation, although there are plans to port ARTISTE to a more open database platform such as mySQL.

These limitations are addressed by extending the protocol in those areas where it does not cover image content-based querying (notably CQL), and by maintaining close contact with the members of the ZING group developing the SRW specifications, both via the ZING mailing list and at various face-to-face meetings. As an early implementer of the draft SRW specification [7], ARTISTE has given feedback to the community on implementation and modification of SRW.

### 5.3.6 ARTISTE Search and Retrieve Web service

ARTISTE supports SRW with a modified version of CQL. The ARTISTE modifications made in the Common ARTISTE Query Language (CAQL) are presented in the next section. A diagram of the ARTISTE SRW architecture is shown in Figure 1.

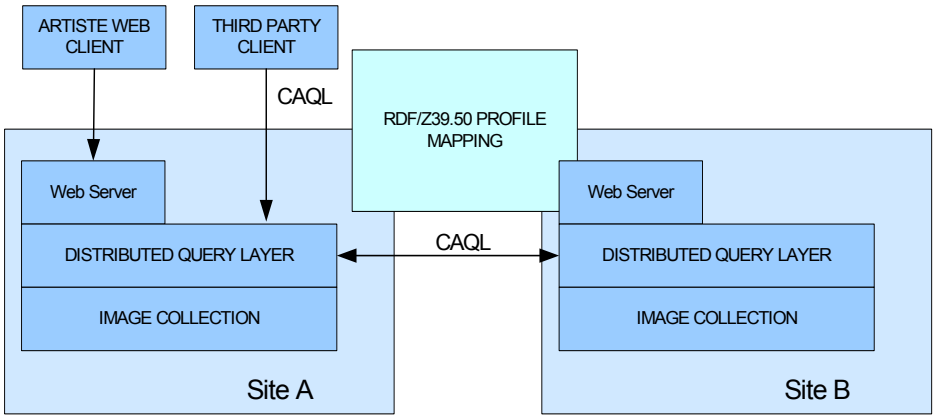


*Figure 22 ARTISTE SRW architecture*

The ARTISTE SRW Server receives a request from a client in the form of an XML message via SOAP/RPC. This request contains a query expressed in the Common ARTISTE Query Language (CAQL), which is based on CQL.

The SRW Service parsers the request parameters and uses the query ontology described in the ARTISTE Core RDF schema [16] to build a corresponding ARTISTE query.

If the SRW request contains an image content-based query, then the client passes a reference to the query image to the server in the form of a URL. The ARTISTE SRW Server then retrieves the query image from the specified web server and passes it, along with the RDF specified query, to the ARTISTE application server.

The ARTISTE application server translates the query to the appropriate SQL and executes it against the database.

The results are returned to the SRW Server which then makes another call on the ARTISTE application server to extract the metadata associated with the returned results in either the default metadata format (Dublin Core) or another format specified in the SRW request.

The SRW Server the responds to the SRW request by sending an XML message containing the URL of image results and RDF encoded metadata back to the client.

The ARTISTE SRW runs on the Tomcat JSP/servlet container engine under Apache.

### 5.3.7 ARTISTE extensions to CQL

This section describes the ARTISTE extensions CQL

The ARTISTE CAQL expands CQL 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 ARTISTE CAQL. The ARTISTE CAQL 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 CAQL queries are given below.

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

This query combines a Dublin Core 'Creator' metadata search with a image content-based query that uses the 'CCV' algorithm [2] to find images that are 'SimilarTo' the referenced query image 'test.jpg'.

```
dc.Subject = TEXTILE and artisteCore.VisibleLightImage part of
MCCV http://artiste.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' algorithm [2] to find images that have the referenced query image as 'part of' them, i.e. as a sub-image.

The attribute/value parts of the queries can be an RDF URI pointing to an element as defined in a schema supported by ARTISTE, e.g:

```
http://artiste.it-innovation.soton.ac.uk/test_rdf/dc.rdf#title
```

Alternatively, the attribute/value parts of the queries can be a z39.50 IndexSet.IndexCode construction using the Bib-1 Attribute Set, e.g:

```
Bib-1.1100

Bib-1.1103
```

In this case, the Bib-1 code is mapped via a Bib-1 RDF schema to the local database schema.

### 5.3.8 Summary of SRW support for ARTISTE functionality

Table 2 shows the ARTISTE functionality that is provided through the ARTISTE SRW.

| ARTISTE metadata and image content based searching functionality | Supported by ARTISTE SRW |
|---|---|
| Metadata Query | Y |
| Image Query | Y |
| Query Across Multiple Databases | Y |
| Query Across Multiple Sites | Y |
| Query Using Multiple Metadata Terms | Y |
| Query Using Metadata Terms & A Query Image | Y |
| Query Using Controlled Metadata Terms | N |
| Query Using Multilingual Metadata Terms | N |

| | |
|---|---|
| Save a Query | N |
| Delete a Query | N |
| Browse Query Runs and Retrieve Results for a Query | N |
| Support dynamics of new & changing metadata | Y |

*Table 2 SRW Support for ARTISTE Functionality*

There are other aspects of ARTISTE metadata and image content-based searching functionality worth describing in the context of supported by the SRW protocol.

Firstly, the user cannot explicitly save or delete queries or browse query runs. However, since the SRW Service returns a result_set_id as part of each response does mean that results for a query can be retrieved and the records iterated over.

Secondly, although a SRW Server typically communicates with a single site or single database, the protocol is designed to allow amalgamation of request to different SRW Servers. Thus a gateway to all the ARTISTE sites could be constructed. This sever-side management of distributed query is the way that ARTISTE currently functions.

Thirdly although the SRW interface does not explicitly expose the multilingual thesauri these are available on the web and users of the SRW are free to consult the RDF document in the creation of their queries.

## 5.4 Summary

ARTISTE provides support for OAI. This allows harvesting of textual metadata from ARTISTE image collections. A possible extension would be to allow image content metadata, e.g. colour distribution or texture, to also be harvested. However, this would require data types for image content to be standardised before support could be added to the protocol.

ARTISTE considers z39.50 as the basis for an open standards based search and retrieval service. However, z39.50 is generally labour intensive to implement and also lacking support for image content-based queries.

SRW is much easier to develop than z39.50 due to use of Web technologies, but still has limitations. Therefore, IT Innovation has extended SRW, in particular by extending the CQL.

IT Innovation has given feedback to the community on implementation and modification of SRW. To do this IT Innovation is releasing an open source client for ARTISTE SRW, a Web browser based client for ARTISTE SRW, a publicly accessible ARTISTE server that supports SRW, documentation on the functionality and use of the ARTISTE SRW and a description of experiences in using SRW as the basis for combined textual metadata and image content-based search and retrieval service. This report is part of the feedback to the ZING community.

Experience of using SRW CQL reveals that manual process has to be used to determine what operators (relation attributes) can be applied to each metadata item (use attribute) for a collection. This involves looking up a description of the use attribute in the Bib-1 attribute set, determining its type from the description in the Bib-1 semantics specification and then determining which relation attributes can be applied by considering the semantics defined for relation attributes in the context of the particular use attribute. The result of this process is an understanding of what use/relation attribute combinations are

theoretically possible. However, this still doesn't say that a particular site that provides a SRW will support these combinations.

In ARTISTE, a different approach is used. Each site publishes RDF that describes the types of all the use attributes and which operators are supported for those types. Therefore, a potential user can determine the limitations of the particular instance of the SRW at that site. Furthermore, since the specification is in RDF, and hence machine readable, this means that the client software could automatically constrain the queries that the user can issue to the SRW. This is a powerful and flexible way for ARTISTE sites to provide a subset of SRW without needing to implement detailed error reporting or issue complex human readable documents that list which parts of the specification are supported.

# 6. Observations

This section presents some observations that go beyond the scope of the individual section summaries or overall conclusions in Section 8.

Digital library support for multimedia content-based search and retrieval is in its infancy and is not accommodated by current standards. At the moment, image library services typically use textual metadata as the basis for collection searching. It is possible to incorporate image retrieval by passing image references from server to client as part of the search results, for example as URLs. Indeed, this approach can allow existing standards such as OAI and SRW to be used.

We believe the full benefit of multimedia search and retrieval can only be realised through seamless integration of content-based analysis techniques. However, not only does introduction of content-based analysis require modification to existing standards as outlined in this report, but it is also requires a review of the use of semantics in achieving digital library interoperability. In particular, machine understandable description of the semantics of textual metadata, multimedia content, and content-based analysis, can provide a foundation for a new generation of flexible and dynamic digital library tools and services. In essence, this is the application of Semantic Web [21] techniques to digital library interoperability.

Overall, we found that RDF was a flexible way to integrate legacy image collections into a search and retrieval service. Integration of the images and metadata is simplified since the underlying collection schema doesn't need modification. Furthermore, if modifications or additions are made to the collection schema, then these can be accommodated through RDF mappings rather than having to redevelop the search and retrieval application. RDF is also suitable for supporting multilingual translation of attributes names and any associated human readable description of their semantics. The use of RDF allows a site to publish descriptions of its search capabilities as well as the metadata and images within the collection. This allows users to dynamically construct queries that are constrained to the available query functionality.

Existing standards do not use explicit semantics to describe query operators or their application to metadata and multimedia content at individual sites. However, dynamically determining what operators and types are supported by a collection is essential to robust and efficient cross-collection searching. Dynamic use of published semantics would allow a collection and any associated content-based analysis to be changed by its owner without breaking conformance to search and retrieval standards. Furthermore, individual sites would not need to publish detailed, human readable descriptions of available functionality.

# 7. Contribution to standards

ARTISTE has both used and attempted to influence standards for metadata and digital libraries. A large part of ARTISTE is concerned with use of existing standards for metadata frameworks, e.g. RDF, and application domain metadata content, e.g. Dublin Core. However, one area where existing standards have not been sufficient is multimedia content-based search and retrieval. A proposal has been made to ZING for additions to SRW. This will enable ARTISTE to make a valuable contribution to this rapidly evolving standard.

In developing the ARTISTE system, IT Innovation has made efforts to disseminate its activities to the digital library community. This has involved presentation at events that include the 10th International Web Conference WWW2002 [22], a European workshop on Digital Libraries organised by DELOS [23], the International Conference on the Challenge of Image and Video Retrieval CIVR2002 [24], and direct participation in ZING workshops. Outside of the technical domain of implementing digital libraries, ARTISTE has been disseminated at many application domain oriented events [25][26][27][28]. A public Artiste demonstrator supporting a Web browser client is hosted by IT Innovation [31] and there is an established ARTISTE user group with 70+ members who receive regular newsletters and have the opportunity to provide feedback on ARTISTE prototypes [32].

ARTISTE is one of the early implementers [9] of SRW and has extended the capabilities of SRW to enable image content and metadata based searches over multiple ARTISTE collections. Having emerged from the digital library community, z39.50 has been traditionally concerned with text based searching. ARTISTE has been working with ZING to incorporate the ability to deal with content-based searching of images and thus expand international standards of information retrieval.

It is only recently that ARTISTE has been in a position to provide significant input to ZING since input is based on experiences in developing a working prototype of extensions to SRW. Some relatively minor contributions have already been made to ZING through informal discussions at conferences and events and through email dialog with members.

A publicly accessible ARTISTE SRW [29] is hosted at IT Innovation and supports access to a combined textual metadata and image content-based retrieval against a small collection of 1000 images from the Victoria and Albert Museum. This report, an open source Java client for the Server, supporting documentation, and previous public presentations together constitute a suggestion for additions to the SRW standard supported by a reference implementation. IT Innovation work will continue in the area of interoperability of digital libraries over the next 3 years, for example as part of the EC funded SCULPTEUR project [30]. This will provide further opportunity for contribution to standards.

# 8. Conclusions

This document has presented the three stage approach ARTISTE has taken to provide an open and flexible solution for combined metadata and image content-based search and retrieval across multiple, distributed image collections.

Firstly, RDF is used to map local metadata schema to common standards such as Dublin Core. This establishes common semantics for content when searching across collections.

Secondly, Query Items (e.g. images, textual metadata attributes), Query Operators (e.g. SimilarTo, Contains, Equals) and the rules for their combination (e.g. SimilarTo can be applied to Images) are all explicitly specified, published and supported in the ARTISTE software. In this way, metadata and image content-based analysis to be combined into a single search queries.

Thirdly, an open interface and protocol for cross-collection multimedia search and retrieval is used that advances open standards for remote access to digital libraries.

The use of RDF mapping provides a flexible solution to cross-collection searching. Mapping to common semantics requires no changes to local metadata and schemas. Multilingual translation of metadata attribute names allows the user to use their native language when specifying which attributes to search over for multiple collections.

RDF schema for metadata, image content, query operators and rules allows combined textual metadata and image content-based querying. This single framework allows both image content and textual metadata queries to be declaratively and seamlessly integrated into a single query language. Furthermore, the capabilities of a particular site can be determined by accessing the published RDF to allow queries to be dynamically constrained to the available query functionality.

As well as supporting OAI, ARTISTE has extended the capabilities of ZING's proposed Search and Retrieve Web service to accommodate image content-based analysis. In particular, enhancements have been made to the ZING Common Query Language to include operators for image content, and image exchange based on URLs. IT Innovation is a member of the ZING implementers group and is participating in ZING to provide feedback on the proposed standards and to encourage support for multimedia aspects.

ARTISTE has both used and provided input to standards for metadata and digital libraries. A proposal has been made to ZING for additions to SRW, supported by documentation, public presentation, hosting of a reference server implementation and release of an open source demonstration client. These will enable ARTISTE to make a useful contribution to this rapidly evolving standard. The current momentum will be maintained into the SCULPTEUR project, which seeks to extend the capabilities of ARTISTE into 3D representations, and has an explicit activity for IT Innovation to develop and disseminate multimedia digital library interoperability protocols.

# 9. Glossary

| | |
|---|---|
| API | Application Programming Interface |
| OAI | Open Archives Initiative |
| ORDBMS | Object Relational Database Management System |
| RDF | Resource Description Framework |
| SRW | Search and Retrieve Web service proposed by ZING |
| URI | Uniform Resource Identifier. A common form of URI is a Web page address. |
| ZING | 'z39.50 into the Next Generation' – initiative to build upon z39.50 |
| z39.50 | Digital library standard for text-based search and retrieval |

# 10. References


[1] ARTISTE EC IST project number 11978 is part funded by the EC under the IST Fifth Framework. http://www.artisteweb.org

[2] Addis, M., Lewis, P., Martinez, K. "ARTISTE image retrieval system puts European galleries in the picture", Cultivate Interactive, issue 7, 11 July 2002 http://www.cultivate-int.org/issue7/artiste/

[3] DublinCore metadata initiative http://www.dublincore.org/

[4] RDF Resource Description Framework http://www.w3.org/RDF/

[5] Open Archives Initiative http://www.openarchives.org/

[6] z39.50 http://lcweb.loc.gov/z3950/agency/

[7] ZING Search and Retrieve Web service http://www.loc.gov/z3950/agency/zing/srwu/srw.html

[8] ZING: z39.50 into the Next Generation http://www.loc.gov/z3950/agency/zing/zing.html

[9] ZING SRW Implementors group http://www.loc.gov/z3950/agency/zing/srwu/implementors.html

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

[11] M.F.A.Fauzi and P.H.Lewis Query by Fax for Content Based Image Retrieval Challenge of Image and Video Retrieval, London, July 2002. Proceedings to be Published in Lecture Notes in Computer Science, Springer Verlag.

[12] Stephen Chan and Kirk Martinez and Paul Lewis and C. Lahanier and J. Stevenson Handling Sub-Image Queries in Content-Based Retrieval of High Resolution Art Images. International Cultural Heritage Informatics Meeting p.157-163. September 2001

[13] RDF Model and Syntax Specification http://www.w3.org/TR/REC-rdf-syntax

[14] RDF Vocabulary Description Language 1.0: RDF Schema http://www.w3.org/TR/rdf-schema

[15] W3C RDF Validation Service http://www.w3.org/RDF/Validator/

[16] Artiste Core RDF schema http://artiste.it-innovation.soton.ac.uk/rdf/ArtisteCore.rdf

[17] Dublin Core Element Set http://www.dublincore.org/documents/dces/

[18] IEFT RFC 1766 Tags for the Identification of Languages http://www.ietf.org/rfc/rfc1766.txt?number=1766

[19] Xmethods BabelFish Web Service http://www.xmethods.com

[20] CERES and National Biological Information Infrastructure (NBII) Biological Resources Division (BRD) Draft RDF Server Standard Documentation http://ceres.ca.gov/thesaurus/RDF.html

[21] The Semantic Web http://www.semanticweb.org/

[22] P. Allen, M. Boniface, P. Lewis, K. Martinez "Interoperability between Multimedia Collections for Content and Metadata-Based Searching", Eleventh International World Wide Web Conference, Honolulu, Hawaii. 7-11 May 2002 http://www2002.org/CDROM/alternate/196/index.html

[23] P Allen, DELOS Russia-EC Digital Library Workshop, Moscow. 8 June 2001

[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] Presentation of the ARTISTE prototype on stand 38, EuroChina 2002, Beijing, 15th-20th April 2002

[26] W. Sterling, G. Presutti, ARTISTE: A Fine Art Repository for Museums with Image Content-based Search Capability, 2001 Multimedia Technology and Applications Conference, University of California, Irvine, California, USA, November 7-9, 2001 (presentation and proceedings paper)

[27] W. Sterling, G. Presutti, The ARTISTE Project: Architecture and Evaluation, 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), Orlando, Florida, USA, July 14-18, 2002 (presentation and proceedings paper)

[28] "Artiste: An integrated Art Analysis and Navigation Environment" presented at EVA 2000, Florence, 31st March 2000 (oral presentation)

[29] ARTISTE Search and Retrieve Web Service demonstration http://artiste.it-innovation.soton.ac.uk/srw

[30] SCULPTEUR EC IST project number 35372 http://www.sculpteurweb.org

[31] ARTISTE demonstration system http://artiste.it-innovation.soton.ac.uk

[32] Artiste User Interest Group, currently a 70 strong mix of academic and industrial members who periodically receive newsletters on the ARTISTE project and provide feedback on ARTISTE prototypes. Contact: Margaret Cecil-Wright, AIUG Co-ordinator, Editor, Artiste Newsletter mcw@it-innovation.soton.ac.uk

[33] AQUARELLE "SHARING CULTURAL HERITAGE THROUGH MULTIMEDIA TELEMATICS" http://aqua.inria.fr/