Capturing e-culture: Metadata in MultiMatch

Neil Ireson - University of Sheffield - n.ireson@dcs.shef.ac.uk
Johan Oomen - Nederlands Instituut voor Beeld en Geluid – joomen@beeldengeluid.nl

This position paper briefly introduces MultiMatch project and the current state of research regarding handing of heterogeneous metadata and approach towards interoperability. Several standards (Dublin Core, FRBR, MPEG-7, CIDOC) have been studied but none of these could meet the requirements defined by the project.

1. The MultiMatch Project

Our shared cultural heritage (CH) is an essential part of our European identity, transcending cultural and language barriers. The aim of the MultiMatch project is to enable users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This will be achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. The development of the MultiMatch search engine can be divided into four areas:

Data Collection
- crawl the Internet to identify websites with CH information, locating relevant texts, images, audio and videos
- likewise identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved

Data Analysis
- automatically classify the results, in a semantic-web compliant fashion, based on document content, metadata, context, and on the occurrence of relevant CH concepts
- automatically extract relevant information which will then be used to create cross-links between related material, such as biographies, exhibitions of work, critical analyses, etc.

Indexing
- organise and further analyse the material crawled to serve focused queries generated from user-formulated information needs

Search and Retrieval
- interact with the user to obtain a more specific definition of information requirements
- organise and display search results in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers

Within the scope of the project 4 languages (Dutch, English, Italian and Spanish) and 4 media types (text, images, audio and video) are considered.

2. Choosing the appropriate metadata representation

The project started to explore how to capture the dimensions of the data by providing an overview of current practice regarding knowledge representation in the cultural heritage domain. As metadata standards enable interoperability between systems and organisations that information can be exchanged and shared, the overview provided the basis for the approach towards interoperability that will be adopted within the MultiMatch project.

This work is documented in the deliverable D2.1: First Analysis of Metadata in the Cultural Heritage Domain. In order to systematically study current practices, the sub-domain definition advocated by the DEN (Digital Heritage Netherlands) and ePSINet (the European Public Sector Information Network[1]) was used. This study included a descriptive overview of the metadata schemas and semantic resources (i.e. thesauri, controlled vocabularies) widely used within the organizations belonging to the specific sub-domains.

A scheme or vocabulary is included only if the following criteria are met:

- it is constructed and maintained by a renowned institute in one of the sub-domains *and*,
- available in electronic form *and*,
- publicly available; in other words, there may be financial but no copyright hindrances to apply them in MultiMatch *and*,
- it is proven an international standard *or* a local standard, in use nationwide.

Forty metadata schemas and semantic resources have been identified and analyzed in a structured fashion. It became clear that the uptake of international established controlled vocabularies is quite limited. Local and nationally established/managed vocabularies are therefore predominant. Part of the reason for this is that the available international controlled vocabularies are still not available in every European language.

The table below lists the most relevant metadata Schema and Controlled vocabularies currently in use in the European cultural heritage sector.

	Schema	**Controlled vocabularies**
Archives	EAD and ISAD(G)	IPTC thesaurus, ISAAR (CPF), Thésaurus architecture et patrimoine, UK Archival Thesaurus
Libraries	FRBR, MARC, MODS and METS	DDC, UDC, LCSH and RAMEAU
Museums	CDWA, Object ID, VRA	AAT, ULAN, TGN
Educational sector	IEEE LOM	ERIC thesaurus
Audiovisual sector	P_META and SMEF-DM	-
Geospatial sector	CSDGM and ISO 19115:2003	-
Generic	CIDOC, DOI, DCMI, MPEG-7/21	URI, RFC1766, ISO3166, …

The methodology from De Sutter (et. al.) in their paper "Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries"[2] was used to select which standard could be used within the MultiMatch project. The following standards were selected for further analysis:

- Dublin Core: because it is in use through the whole of the cultural heritage domain.
- MPEG-7: because it can handle multimedia in a way appropriate for MultiMatch.
- FRBR: because it provides a data model with relationships and a hierarchy that are probably useful for MultiMatch. (Annex 3 includes the graphical representation of the FRBR entity-relationship model).

[1] http://www.epsigate.org/

2 Sutter, R de. [et. al.] Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries. Published in: Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, Rafael C. Carrasco (Eds.): Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings. Lecture Notes in Computer Science 4172 Springer 2006.

- CIDOC CRM: because it provides a reference model for the cultural heritage domain. (Annex 4 includes the graphical representation CIDOC class hierarchy).

MultiMatch will handle various types of metadata:

1. Descriptive metadata – both metadata that formally describe the object (for example title, creator, creation date) as well as some semantic elements (for example subject keywords, geographic places);
2. Technical metadata – probably mainly concerning the surrogate or the image of the cultural heritage object, and less concerning the physical cultural heritage object itself;
3. Administrative metadata – some metadata to administer the objects concerned (e.g. content provider name, location information, language of record, record number), possibly also some metadata for the rights management. For instance, the extent to which metadata on copyrights are needed within the central metadata schema is at this point not clear.
4. Examples of typical metadata that will probably not be the focus of the MultiMatch metadata schema include administrative and technical data on museum objects that are needed for the internal management of the museum collection (typically gallery and museum information systems)

The next step was to choose/built upon these standards and design the appropriate metadata representation approach for the project. The main factors that influenced this work can be divided into three areas;

- Meeting the specification of the user requirements,
- Representing the concepts which are present in the data and
- Interoperability: mapping from content provider legacy metadata and to current "standard" metadata schemas.

2.1 Meeting the specification of the user requirements

The MultiMatch project is following a user-centred design strategy and so a good deal of the initial effort has been directed at interviewing the end-users of the system to determine their requirements. These user requirements are translated into a functional specification which is used to guide the development of MultiMatch so that the final system will meet the needs (and expectations) of the end-users. Obviously there is not a singular concept of an end-user, and different users having different levels of expertise and information search needs will present different requirements which should be considered in the system development.

In terms of the influence on the metadata, the user requirements give and indication as to the concepts which are required by the user and the relative importance of those concepts.

Many of the findings of the user requirements are, as one would expect, intuitive, such as the fact text and images are the primarily important media types, and that creations and their works are the main concepts in search, with typical queries involving: proper names, places, titles, general subjects. There was also less definite requirements, such as clearer semantics in faceted browsing and clustering so that the categorisation (clustering) of the search results "make sense". This strongly points to the need to use standard (and well defined) vocabularies and subject taxonomies.

Given that in MultiMatch all search facilities will be **translingual**, i.e. the user will formulate queries in a given language and retrieve results in all languages covered by the prototype. Users

expressed a strong desire to have control over the process, i.e. selection of the appropriate translation in the case of a word with multiple meanings and the ability of not translating certain expressions.

The notion of a controllable and transparent search process was seen as a general requirement on the system, a key reason for this is the ability to determine the “authoritativeness” of any information which results from a user search.

From the expert users survey we can conclude that, on average, experts tend to classify searches for information about creators (authors) and creations (works of art and masterpieces) as their most common search tasks. Therefore, in MultiMatch we have initially decided to focus two types of specialized searches on creators and creations, although specialized searches focused on other relevant categories will also be considered.

2.2 Representing the concepts which are present in the data

Whilst it is crucial to provide a representation to fulfil the user requirements, these will not fully specify the requirements of all users; as only a selection are interviewed, needs are dynamically changing and users concepts of what they want are sculpted by their experience of what is available. In addition, by examining the data it is possible to determine the issues which arise when representing the concepts relating to cultural heritage objects.

Therefore the metadata should consider the ability to adequately represent the concepts which are extractable from the data itself, so that concepts (including possibly unforeseen concepts) can be suitably represented. In the MultiMatch project this data includes direct representations of Cultural Heritage objects (images, texts, etc.) and data describing those objects, which is presented in a human rather than machine readable format.

Within MultiMatch there is an obviously need to represent the creator(s) of an art work; this can include information such as: names, birth/death place and time, colleagues, etc. However when examining the data it is seen that values can; be ambiguous (an individual can be known by several names), have varying degrees of precision (i.e. country, province, town) or may be contested (such as whether the artist actually created the work). Also associations can have varying degrees (artists may collaborate over many years or merely on a single art work).

The ambiguity, imprecision and uncertainty of data is accentuated by the multilingual nature of the data and fundamentally by the use of automatic techniques to extract information form the data. Where a concept is seen to have conflicting values this could be due to a genuine difference of opinion in the CH domain, an error in the data or an error in the extraction process.

One of the key features which was highlighted by the user requirement analysis is the need to represent the authoritativeness of the information presented to the user. It is therefore important for the MultiMatch representation to be able to express the ambiguous, imprecise and uncertain nature of the information to the user. Also it is potentially useful/necessary to provide an “audit-trail” to the source(s) and process(es) which have been used to acquire the information.

In addition to the MultiMatch project extracting metadata from textual data, it also extracts metadata from other media typesaudio, still images and video. Although to an extent audio (transcripts) can be seen as noisy text and video is seen as (keyframe) images and audio (text).

The metadata which is extracted from images represents the underlying, low-level features of an image however what the user requires is the linking of this to a meaningful (semantic) representation of the image. Thus the metadata format must cope with low-level and high-level concepts and the links between these representations.

2.3 Metadata Interoperability

Interoperability is concerned with the capability of different information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information.

Cultural heritage organisations employ professional cataloguers to annotate the objects in their collections. A metadata model and cataloguing rules formalises the way objects are described in the catalogue. Thesauri are often the basis to assign keywords to objects. The situation gets more complex if the goal is to create connections between objects and between collections, even more so if this process needs to be automated to the largest extent possible. It involves matters of syntactic (i.e. schema-level) and semantic interoperability; where both are essential.

Semantic interoperability is characterised by the capability of different information systems to communicate information consistent with the intended meaning of the encoded information (as intended by the creators or maintainers of the information system). It involves processing of the shared information so that it is consistent with the intended meaning and encoding of queries and presentation of information so that it conforms to the intended meaning regardless of the source of information.

So, the basic interoperability questions MultiMatch has to deal with are related to:

- Automatic extraction of metadata field. Will the MultiMatch technology, which we are going to build, be able to generate the metadata fields that are in the MultiMatch metadata schema?
- Mapping of original metadata into the MultiMatch schema. The project will have metadata already available in the content providers databases to be integrated with the MultiMatch schema. Is this mapping feasible?
- Support semantic interoperability. As MultiMatch covers a wide domain, this is a big challenge. How will this be reached? How are multiple languages supported?
- Role of ontologies. Generic metadata schema, like Dublin Core, reveal the existence of overarching concepts. This might be the basis of creating an ontology to create interoperability between individual schema.

In the research conducted by MultiMatch, it became clear that next to the Dublin Core element set, also the entity-relationship model FRBR and CIDOC conceptual reference model can prove valuable solutions to provide the level of interoperability required.

3. Conclusion

The factors influencing the design of the MultiMatch metadata schema come (“top-down”) from requirements of the user, (“bottom-up”) from the data and information extraction process and from the need to be interoperable within the cultural heritage domain. There is no single “standard schema” which meets all these requirements. In terms of interoperability Dublin Core

(DC) is the most widely used standard, and whatever metadata schema is implemented the relevant parts must be mapped on the DC. A number of standard refinements of DC are also available (DCMI, VRA, etc.) and providing the ability to map to these increases the information which can be transferred. The area of metadata for low-level multimedia features is an evolving one although the MPEG-7/21 descriptors seem to offer a powerful (and increasingly standard) way of representing and communicating such information. Finally there are reference models such as CIDOC CRM and FRBR which offer an over-arching representation of the CH domain.

In terms of semantic interoperability the desire is to provide a means which allows for flexibility and coverage (providing information to the widest possible user base) and expressiveness (is as informative as possible). Currently mapping to and from the Dublin Core provides the widest coverage however much of the richness of information which is within the MultiMatch metadata will be lost, thus it is necessary to provide the metadata in more expressive representations (MPEG-7, VRA, CIDOC) allowing the receiver to utilise such information, obviously ensuring that the semantics behind the metadata schema are clearly stated and adhered to during the process of populating the metadata.