MultiMatch: Multilingual / Multimedia Access to Cultural Heritage

Giuseppe Amato, Franca Debole, Carol Peters, Pasquale Savino

Institute of Information Science and Technologies - CNR
Pisa, Italy
{Franca.Debole}@isti.cnr.it

Abstract. 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 has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. MultiMatch aims at complex, heterogeneous digital object retrieval and presentation.

1 INTRODUCTION

Online Cultural Heritage (CH) content is being produced in many countries by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events in Europe or Asia. In order to gain a full understanding of such events, including details contained in different collections and exploring different cultural perspectives requires effective multilingual search technologies. The EU FP6 MultiMatch project is concerned with information access for multimedia and multilingual content for a range of European languages. MultiMatch tries to offer "complex object retrieval" through a combination of focused crawling, and semantic enrichment that exploits the vast amounts of metadata available in the cultural heritage domain. The MultiMatch search engine was developed with specialised search facilities for multilingual access to cultural heritage material in diverse media. The aim was to present the user with the "detailed picture" of complex CH objects. The overall goal of the project was to build a fully operational system prototype, designed and refined according to the requirements of diverse user classes. Users can search information using their preferred language, searching for all types of digital objects, accessing only the sites that contain information potentially relevant to their request, retrieving mainly relevant items, and viewing the query results in an organized structured fashion. Standard and

ontology-based descriptions of content are used, thus providing an interoperable semantic framework for intelligent multimedia object delivery. Metadata automatically extracted from CH material was mapped onto this framework. The system has been demonstrated for the main languages of the cultural heritage institutions in the consortium: Dutch, Italian, Spanish, English and also German and Polish, but it was extendible to other languages.

2 MOTIVATION

Europe's vast collections of unique and exciting cultural content are an important asset of our society. On the web, cultural heritage (CH) content is everywhere, in traditional environments such as libraries, museums, galleries and audiovisual archives, but also reviews in popular magazines and newspapers, in multiple languages and multiple media.CH objects on the web are no longer isolated objects, but situated, richly connected entities, equipped with very heterogeneous metadata, and with information from a broad spectrum of sources, some with authoritative views and some with highly personal views. 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 has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. The main source of information stored in the MultiMatch prototype system is composed of cultural heritage objects obtained through crawling and indexing of material obtained from cultural heritage sites, web encyclopedias (e.g. Wikipedia), digital libraries of specific cultural heritage organizations, OAI compliant digital resources, and RSS feeds from cultural web sites. The cultural heritage search and navigation facilities envisaged by MultiMatch cater for these information needs by presenting users with a composite picture of complex CH objects. For instance, in reply to a user's request for information on Van Gogh, the MultiMatch engine can present information on Van Gogh from multiple museums around Europe, in multiple languages; it could complement this with pointers to Van Gogh's contemporaries, with links to exhibitions on Van Gogh, to reviews of these exhibitions, to blog entries by visitors to these exhibitions, and to background information taken from online resources or dedicated sites. The MultiMatch search engine has been developed with specialised search facilities for multilingual access to cultural heritage material in diverse media.

3 THE SYSTEM

The MultiMatch search engine is able to:

– identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved;

- crawl the Internet to identify websites with CH information, locating relevant texts, images and videos, regardless of the source and target languages used to write the query and/or describe the results;
- automatically classify the results on the basis of a document's content, its metadata, its 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 the biography of an artist, exhibitions of his/her work, critical analysis, etc.;
- organise and further analyse the material crawled to serve focused queries generated from information needs formulated by the user;
- interact with the user to obtain a more specific definition of initial information requirements;
- the search results are organized in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers.

The MultiMatch search engine enables the user to retrieve cultural objects through different modalities:

1. The simplest one is a traditional free text search. This search mode is similar to that provided by general purpose search engines, such as Google, but MultiMatch provides more precise results and with support for multilingual searches (English, Italian, Spanish, Dutch, German, and Polish). Multilingual searches are performed either through machine translation or by using a general purpose dictionary extended to include terms which are CH specific.
2. Multimedia searches, based on similarity matching and on automatic information extraction techniques.
3. Metadata based searches.
4. A browsing capability allows users to navigate the MultiMatch collection using, among others, a web directory-like structure based on the MultiMatch ontology.

Searches can be made at three main levels of interaction: (a) Default search mode, (b) Specialized search mode, (c) Composite search mode. The *default* search level is provided for generic users. In this way, given a general query, MultiMatch retrieves all the cultural objects, web pages and multimedia content that best suit the query. Merging, ranking and classification of these results are also performed. Users with a more precise knowledge of system functionality, and with specific search needs, may use one of the specialized interaction levels available. These allow the user to query specific search services. In this way, MultiMatch includes standalone image, video and metadata-based searches, each with its own search fields, display and refinement options. It also includes a set of browsing capabilities to explore MultiMatch content. The "composite search mode" supports queries where multiple elements can be combined. For example, it is possible to search using the metadata fields associated with each document, but combining this restriction with free text and/or image similarity searches. In the following subsections we describe the MultiMatch approach to support the specialized search:

- creators search. The general idea of this specialized search level is that, for a given type of cultural entity (creator/creation), the user can query the MultiMatch system to retrieve all the information available about it (e.g. the user can query about Van Gogh and then retrieve all the information available about the painter). MultiMatch creates relations between cultural objects to allow the user to browse and discover information related with his current search.
- audiovisual search. This type of information can be considered as multimodal, which implies that pure visual contents (images and videos) are also related with spoken contents, associated metadata, and texts describing the contents. The MultiMatch system provides three different specialized searches on multimedia contents:
  - image search. MultiMatch offers the possibility of retrieving still images and video keyframes based on text and image queries using multimodal searching.
  - video search. MultiMatch offers the possibility to search for video contents using text queries and also image queries.
  - audio search. Users is able to perform audio search to retrieve audio documents and also video documents by way of their speech tracks. An index built from these transcripts makes audio search possible. The user submits a free text query. Audiovisual documents relevant to this query (i.e. containing the query in the speech recognition transcript) is shown, allowing the user to start playing the audio or video document before the occurrence of the first query word.

4 CONCLUSION

The project was completed on October 31st, 2008 and all planned objectives were achieved. The MultiMatch project involved 11 partners: Istituto di Scienza e Tecnologie dell'Informazione - Consiglio Nazionale delle Ricerche (ISTI-CNR), University of Sheffield, Dublin City University, University of Amsterdam, University of Geneva, Universidad Nacional de Educacion a Distancia, Fratelli Alinari Istituto Edizioni Artistiche SpA, Netherland Institute for Sound and Vision, Biblioteca Virtual Miguel de Cervantes, OCLC PICA, WIND Telecomunicazioni SpA. The project website is active (http://www.multimatch.org) and an online demo of the MultiMatch prototype can be accessed by registered users (free registration available).

Acknowledgement
Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST- 033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.