MultiMatch: Multilingual / Multimedia Access to Cultural Heritage

Carol Peters & Pasquale Savino
ISTI-CNR, Pisa, Italy
MultiMatch: The Initial Idea

Problem:
- The Web contains a wealth of fragmented CH information but users are left to discover, interpret and aggregate it.

Objectives:
- Develop a search engine that provides targeted, enriched access to heterogeneous CH objects
  - across all media types and language boundaries
  - supporting various user classes
  - with aggregate views on complex task scenarios
- Assist CH institutions to raise visibility and disseminate content
The Project

- 6FP Project
- Technology-enhanced Learning and Access to Cultural Heritage
- Started: May 2006
- First prototype: August 2007
- Second Prototype: July 2008
- Evaluation and field trials: October 2008
The Consortium

Academia
- Istituto di Scienza e Tecnologie dell’Informazione
- University of Sheffield
- Dublin City University
- University of Amsterdam
- University of Geneva
- Universidad Nacional de Educación a Distancia

Industry
- OCLC
- WIND Telecomunicazioni S.p.A.

Cultural Heritage
- Fratelli Alinari Istituto Edizioni Artistiche SpA
- Netherlands Institute for Sound and Vision
- University of Alicante – Biblioteca Virtual Miguel de Cervantes
Main Research Challenges

From document to complex objects retrieval

- **Focused crawling** for acquisition of CH-related information from heterogeneous MM resources
- CH concept and relation extraction using information extraction and text mining techniques
- **Multimedia search** and mixed media search
- **Multilingual management** with support for query formulation, cross-language retrieval and summarization
- Integration and representation of related objects
- Presentation of aggregated search results
- **User support** (e.g. search history, annotation facilities, personalized presentation of results, etc.)
Selection and adaptation of content

- Objective of this activity is to collect and prepare a significant amount of content to test and validate the first system prototype
- Selection of content from MultiMatch content providers’ archives
- Selection of content from a preliminary list of relevant CH institutions web sites (white list crawl)
- Selected set of wikipedia pages
- Audio corpus selected from web
- OAI compliant records from TEL
Metadata modelling and interoperability

- Definition of a MultiMatch conceptual reference framework suitable for Cultural Heritage
  - Metadata schema
  - Thesauri for artists names and descriptions, geospatial information
- Simple to use (interoperability)
- Suitable to the application of automatic population techniques
Indexing and information extraction

- Automatic extraction of indexing features for all media (text, speech, images, video) and crawled data
- Text indexing for four languages (Italian, Spanish, Dutch, and English, plus others)
- Automatic generation of inter-document links
- Development of algorithms for classification and information extraction
  - Creators, type/genre, subject, place/time, art objects/works
- Semantic enrichment of documents
- Linking information sources
  - Thesauri from outside the system
  - Documents within the system
Multilingual support

- Provide system with monolingual and multilingual search functionalities (initially four languages, to be extended)
- Provide effective translation strategies e.g. multilingual dictionaries, machine translation, thesaurus term expansion
- Multilingual query expansion
- Dynamic summarization
Machine Translation

- WorldLingo commercial machine translation system used under licence
- Supports all 12 language pairs for the four selected languages
- Easy to use and integrate into prototype
- Good well documented API
Dictionary-based Query Translation

- Word-by-word (phrase-by-phrase) translation via a bilingual dictionary look-up
- Translation resources
  - General translation lexicon
    FREELANG dictionary www.freelang.net
    Universal dictionary www.dicts.info
  - Domain-specific translation lexicon
    Wikipedia www.wikipedia.org
Domain-specific Translation Example
Multimedia search

- Similarity search based on visual features (low level and high level – faces, objects, etc.)
- Efficient retrieval
- Support of relevance feedback and interactive search
- Combined text and visual search
User interaction and interface design

- User-centred design process
  - Evaluate and refine the interface based on empirical evaluation and usability testing
- Interface supporting multilingual and multimedia search
- Default search on all types of content
- Specialized search on metadata fields and on different media
- Use of semantic structures for search and browse
Project achievements

- **First project phase completed**
  - User requirements analysis
  - MultiMatch conceptual reference framework for Cultural Heritage
  - First prototype of the MultiMatch search engine
    - Baseline for experimentation, user feedback and second prototype development
    - Limited content
    - Based on manually associated semantic concepts
    - Multimedia search – image similarity search; intra-video shot detection
    - Multilingual interaction (4 core languages)
Research issues for second project phase

- Automatic document labeling, enrichment and linking and automatic document classification
- Combination of cross media query results
- Translation and summarization of results
- Innovative User Interface
  - Advanced Multimedia search and browse
  - Automatic and interactive translation covering six languages
  - Optimized search and browse based on semantic structures
Conclusions

- Further evolutions of MultiMatch
  - Use of MultiMatch technologies to build large scale Digital Libraries and a large scale search engine specialized for Cultural Heritage
    » Enlarge the content base through the access to a complete set of CH sites and crawling of a significant part of the web
    » Enlarge the number of languages managed (possibly to cover all EU languages)
    » Invest on system efficiency, system scalability, and robustness
Further information

- MultiMatch Web site
  - http://www.multimatch.eu

- Project coordination: Pasquale Savino
  - pasquale.savino@isti.cnr.it

- Technical coordination: Giuseppe Amato
  - giuseppe.amato@isti.cnr.it

- User group coordination: Sam Minelli
  - sam@alinari.it
Relations between authors

van Gogh Network

Yukio Mishima
Suicide
Depressive

Rembrandt

Paul Cezanne

Henri Rousseau

van Gogh

dutch, flowers, post-impressionism, suicide, sunflowers, van Gogh museum
Relations between Author and works of art

<table>
<thead>
<tr>
<th>Artwork Name</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starry Night</td>
<td>1889</td>
</tr>
<tr>
<td>Cafe at Night</td>
<td>1889</td>
</tr>
<tr>
<td>Sunflowers</td>
<td>1889</td>
</tr>
<tr>
<td>Starry Night Over the Rhone</td>
<td>1889</td>
</tr>
<tr>
<td>Self Portrait with Grey Hat</td>
<td>1889</td>
</tr>
</tbody>
</table>
Specialized image search
Multilingual Search (1)
Multilingual Search (2)
Multilingual Search (3)
Multilingual Search (4)