Gathering requirements for multilingual search of audiovisual material in cultural heritage[1]

Sam H. Minelli (Alinari IDEA, IT), **Jennifer Marlow** (University of Sheffield, UK), **Paul Clough** (University of Sheffield, UK), **Juan Manuel Cigarran Recuero** (Universidad Nacional de Educación a Distancia, ES), **Julio Gonzalo** (Universidad Nacional de Educación a Distancia, ES), **Johan Oomen,** (Netherlands Institute for B&G, NL), **Domenico Loschiavo** (Alinari IDEA, IT)

***Abstract*— The MultiMatch project plans to develop a multilingual search engine specifically designed for access, organisation and personalised presentation of cultural heritage information. This article presents the MultiMatch user requirements analysis methodology, which provided input for the definition of the functional specifications of the system prototype. A description of the potential MultiMatch user is given and the methodology used to identify the user requirements is defined. During the MultiMatch project we used both direct and indirect approaches: we involved the users to derive their requirements (which then informed proposed technical solutions) and also focused on literature and user scenarios derived from long term forecasts and visions. In practice, interviews were carried out to establish what target users needed and expected from cultural heritage search, query logs from cultural content holders were analysed to highlight typical user behaviour, typical usage was depicted though narrative scenarios and possibilities for future long-term needs, and finally past literature and prior experience was used to validate and consolidate the results.**

***Index Terms*—MultiMatch, cultural heritage (CH), education, results aggregation, user requirements.**

I. Introduction

In order to create a system that meets the needs of its target audience, the MultiMatch[2] project has adopted a user-centered design methodology. As an initial step in this process, one hundred person-to-person interviews were conducted with domain experts (educational, tourism, and cultural heritage professionals) in order to collect their opinions and needs. The interviews were conducted mainly face-to-face using a questionnaire, on which a set of scenarios and a vision document were created in order to give the respondents an idea of the functionality of the proposed system. In addition to the interviews, initial analysis of log files from the WIND[3] portal, Alinari IDEA[4] photographic archive business site, Tate Online[5], Biblioteca Virtual Miguel de Cervantes[6] and Sound and Vision[7] search engines were examined, along with the results of previous user studies in the cultural heritage domain. Many potential requirements were initially identified and analysed, the final group being a subset of those requirements deemed to represent the major needs of the user groups studied and matching the project vision. The output of this analysis has been used to generate the functional specification for the MultiMatch system and tailor specific functionality/services.

II. The MultiMatch Vision

Often for individuals seeking access to material related to cultural heritage, the information needed can be found on the Internet. However, it may only be accessible if the information seeker is able to cross various boundaries. One important boundary is that of language. Important information could be available to users, but inaccessible because it is expressed in a language which differs from that normally employed by the user in his/her search. The MultiMatch system aims to deal with this by allowing the user to search and browse content in their preferred language. The user's information need (typically expressed as a query) will activate a multilingual search, gathering the results, organising and presenting them in such a way that the user is able to effectively use them.

Two key ideas underlying MultiMatch are: **multiplicity** and **aggregation**.

Multiplicity: MultiMatch will display multimedia results to the user in multiple languages, with various options for searching/browsing, and with multiple links between pages and sites. Users will be able to pose queries in their preferred language(s) and retrieve material in all languages handled by the project *(i.e. if the search terms are 'ritratto di Giacometti e scultura eseguita da lui,' the system will also execute queries in English ('Giacometti's portrait and a sculpture made by him') and/or other languages*). According to the user's language profile, results in unknown languages will be returned in a way that is interpretable by the user, e.g. with a summary or associated keywords in the user's preferred language, or even with a translation acquired from an on-line machine translation service.

[1] This material is partially based upon work supported by the European Community in the MultiMatch (Multilingual/Multimedia Access to Cultural Heritage, FP6-2005-IST-5, no. 033104) project. This paper is the view of the authors but not necessarily the view of the community.

[2] http://www.multimatch.eu

[3] http://www.libero.it

[4] http://business.alinari.it

[5] http://www.tate.org.uk

[6] http://www. dlsi.ua.es

[7] http://www.beeldengeluid.nl

Users and customers will be able to search text (‘*find critical texts on Giacometti and his period*’), images (‘*Giacometti’s portraits and his sculptures’*), audio and video (‘*Documentary about the life of Giacometti and the places where he lived’*): image search facilities will include text and content-oriented matching; audio and video search will include the capacity to search transcribed speech (at least in English in the project prototype).

Aggregation: The system will aggregate results from diverse sources, and depending on the type of query, can include the results of one or more MultiMatch specialized searches. Specialized search services will be activated to interact with the user to retrieve optimised search results: if the user is looking for images, the specialized search interface will make it possible to further filter the search by collection, by visual criteria, or by metadata-related information.

III. MOBILE APPLICATIONS

One of the most challenging future application of any search engine, and thus also of MultiMatch, is to be available through mobile devices. This enables travel-related questions for cultural heritage topics, e.g. “I’m visiting Florence for a meeting, I’ll have 3 hours free; where is the nearest museum and what exhibitions are featured there?” Furthermore, with improved cross-language searching functionalities, the new generation of search engines could act as a type of travel dictionary. Whilst Google and Yahoo! already provide mobile search services, they generate broad results that require further filtering. The aim of MultiMatch is to address more specifically the needs of cultural users who would gain an added-value service from mobile: multilingual summarization of results.

Query by SMS

A question can be submitted to Google[8] with a mobile SMS (Short Message Service) query: the user sends the query as text messages over the mobile phone or device, which returns answers (without links or web pages; only text) to the questions. Example applications include:

- To get local business listings when the user is on the road and wants to find a place (hotels, restaurants, museums, etc.).
- To get quick answers to straightforward questions.
- To look up dictionary definitions to expand personal vocabulary or prove a point (as example provided from Wikipedia’s resources).
- To get cultural information about the place in which you are (a square, etc.).

Of the functionalities provided by Google, those pertaining to MultiMatch include:

- To look-up dictionary definitions and knowledge as examples provided from Wikipedia’s resources.
- To get cultural information about the place in which you are located (e.g. a square).
- To get preview information about ongoing exhibitions and press releases.

Query by using i-Mode

Google also provides i-MODE search services (text and images) for web and image search. The customer needs to have a data plan for his/her device in order to use these services. Coverage may also differ depending on the customer’s carrier and mobile phone. Alinari[9] has released i-Mode access to its contents and would benefit from a MultiMatch engine focused on cultural heritage content (museums, fine arts, etc.), providing authoritative access to its cultural information. Mobile access to content is, in the long term, so important that many similar services are going to be released (even if at this time the mobile devices do not satisfy the high quality expectations of end users). We are keen to consider future improvements of the devices which will allow the delivery of higher quality content.

IV. INITIAL STEPS: DEFINITION OF THE USERS

Identifying the appropriate user groups and individuals is a key step when defining system requirements. Cultural heritage partners in MultiMatch have thus worked at identifying the most relevant user groups likely to benefit from the services that will be offered by the MultiMatch system. Different classes of user have been identified in the educational, tourism, and cultural heritage sectors. Analyses of target users’ needs and the tasks they perform were performed in order to create scenarios demonstrating the ways in which MultiMatch can be expected to operate for these users.

V. METHODOLOGY

The methodology implemented for MultiMatch draws on currently accepted procedures and in particular, a similar methodology for requirement collection which has been applied by Alinari during the design of a professional interface (aceMedia[10] project) for searching digital contents (in particular: video and images) by using new generation mobile devices.

Current practice suggests that the results from using different data collection methods should be combined to create a more complete set of requirements, in a process known as triangulation (see [6]: pp 93, [14]: pp 317-337). The chosen methodology involved collecting information via a variety of approaches and then triangulating this information to obtain a holistic view of user needs. The means of data collection included 1) the use of past literature and partner experience, 2) an analysis of pre-existing and “competitor” websites (see [3]: pp 260-276), 3) the execution of user interviews to establish expectations and user needs (see [14]: pp 214) and 4) an examination of query log files to identify user search behaviour on a larger scale. This also led to 5) the establishment of narrative scenarios to help identify possible long-term needs.

[8] http://www.google.com/sms/

[9] http://i-mode.alinari.it/home.cgi

[10] http://www.acemedia.org

The user requirements analysis performed by MultiMatch is based both on previous experience acquired by CH institutions involved in MultiMatch (Alinari IDEA and Sound and Vision), and also from past literature. The goal has been to identify the target users and their needs within a predefined and specific context and to map, where applicable, these requirements to features which should be offered by MultiMatch.

This study mainly addressed the needs of users that access cultural heritage information in a professional setting. The motivation is that this kind of user already has well-identified requirements and has had experience in trying to satisfy them with the currently available tools. The analysis has aimed at addressing questions such as what users in the cultural heritage domain typically do on a day-to-day basis (i.e. their work tasks), what type of information they need, and how they look for it (i.e. their search behaviour), what these users would require from an information system like MultiMatch to enable them to carry out their activities more effectively (i.e. functionality), and how these users would expect MultiMatch to respond to their search requests (i.e. presentation).

It was intended to investigate the needs of the "casual" or non-professional consumer of CH information for the purposes of personal interest, entertainment or travel via online questionnaires. The results of the questionnaires have been analysed and will provide feedback to the functional specifications for the second and final project prototype.

VI. LIMITATIONS OF THE METHODOLOGY

Of course the methodology is not free from problems. One of the major difficulties that we face is the fact that we are planning to provide substantially new, research-oriented search functionalities. Typical users, however, tend to express their search needs in terms of what current search technologies offer to them, and are often unable to anticipate new search modes. Our questionnaires overcome this problem by posing questions where new functionalities are defined, and then they are asked to state whether such functionalities could be useful to them, to what degree, and which ones should be prioritized. The counterpart is that then we get questionnaire results which are somewhat biased by what we intend to implement beforehand. It is difficult to completely solve this problem; in any case, we certainly need at least another round of interviews once these functionalities can be demoed.

VII. REQUIREMENTS ELICITATION AND DATA COLLECTION

Deciding on which data collection techniques to use is often difficult and depends on a number of factors, e.g. whether the information required is qualitative or quantitative, the current stage of the project lifecycle (i.e. at the beginning it is likely that there are fewer questions so it may be better to explore issues with interviews rather than questionnaires), available resources, access to stakeholders in the project, the nature of the data gathering technique, the task to be studied and the type of information required.

The various approaches taken to gather requirements in the present study will now be discussed in greater depth.

- **Competitor analysis**: this approach is typically used as part of a business modelling phase. The aim of competitor analysis is to compare industries and features (see [3]: pp 260-276). It provides a snapshot of a marketplace from a customer's viewpoint including the services and features offered by companies. For example, when redesigning websites, competitor analysis is often performed as part of the web design process. The goal of the analysis is to evaluate the features, technology, content, usability and overall effectiveness of services available to customers or users within a domain.
- **Interviews** involve asking someone a set of questions and can be held face-to-face or over the phone. They are good for exploring issues, as questions can be guided and clarified by the interviewer. Scenarios can be used in interviews to get people to describe their day-to-day activities, but a more accurate approach in this regard is naturalistic observation (see below). Interviews mainly produce qualitative data, but some quantitative data can be generated. However, this approach is time-consuming and needs training activities not to influence the interviewees.
- **Analysis of log files**: the operations carried out by existing operational systems are often captured in log files (e.g. transactions carried out by a Web server). These can be examined to provide information that complements other methods of data collection. However, this method only explains what has happened and cannot explain *why* something has occurred.
- A **scenario** is an "informal narrative description" of human activities or tasks in a story. This is a natural way for people to describe their tasks and typically does not include information about particular systems or technologies to support the task. These descriptions can then be analysed to extract requirements (and also to build up models of the domain, e.g. building class diagrams based on extracting nouns from the narrative).

There are also a variety of other methods that could be employed depending on the type of information sought, including focus groups and workshops, naturalistic observation, and studying documentation. These can all yield different insights and perspectives, and elements of these were incorporated into speaking with the users (for example, part of the interviews involved observing how users performed typical tasks).

VIII. RESULTS

The competitor analysis resulted in a list of the most common functionalities employed on the various related cultural heritage sites (see Table 1). Most common among these were the possibility of conducting a free text search and of browsing by category. Roughly 1/3 of the 56 sites surveyed had some sort of multilingual offering, but only two of these offered query translation services. Looking at pre-existing

sites can give ideas for useful features to include in a new system, and can thus reveal areas in which novelty and innovation can occur.

Table 1: Relative proportions of functionalities offered by 56 cultural heritage websites

Functionality	Percent	Example
Free text search	91%	-
Browse by category	71%	www.archinform.net
Advanced search	70%	-
News/Calendar	61%	www.tate.org.uk
Registration/login	45%	-
Multilingual	34%	www.louvre.fr
Geographical search / Map	29%	http://whc.unesco.org/en/map
Shopping	29%	-
Search within results / See "more like this"	29%	www.fotolia.com
Ability to segregate multimedia results by type (if applicable)	29%	www.archive.org
Timeline / Search by time	21%	www.birth-of-tv.org
Change results layout (order by..)	21%	www.artandarchitecture.co.uk
Hierarchical browse	20%	http://www.staffspasttrack.org.uk/
Sitemap	20%	-
Controlled vocabulary	9%	www.tate.org.uk
Colour/layout search	7%	www.hermitagemuseum.org
Query translation	5%	www.fotolia.com
Faceted browsing	3%	http://orange.sims.berkeley.edu/[11]

With regards to information extracted as a result of the interviews, a large set of requirements (hundreds) were initially created and analysed in order to identify: (a) the most requested functionalities (and thus could be considered high-priority), and (b) those requirements that best matched the project objectives and vision. In summary, the main requirements for Cultural Heritage (CH) professionals were that:

- They *do* use the internet widely and as part of their daily work routine, but currently depend largely on generic search engines to find the information they need. A specialized search engine for CH would be beneficial.
- They want to query using both free-text search (natural language) and familiar Boolean operators.
- They would like full capabilities for multimedia retrieval (i.e. images and video as well as text), but in most cases are only accustomed to performing text searches.
- Their main focus appears to be on works of art (creations) and their creator. They also want access to associated information, such as critical reviews, information on exhibitions, different versions of same document.
- They tend to be frustrated by the volume of information available on the same subject and would find information filtering, clustering and aggregation functionalities very useful.
- They demand high precision of results and need to know the source and level of authority of the CH material.
- They need to be able to save both queries and results for future processing and reuse.
- They tend to restrict their searches to their own language, plus English, thus missing information only available in other languages.
- If multilingual search was available, they would like to have the results associated with descriptive snippets in their own language (preferably) or English (optionally).
- If the information collected is not clear and meaningful enough then they would not select the link even if the it has high ranking.

The log file analysis focused on examining the most popular search queries submitted to the various cultural heritage sites mentioned previously. As a result, it was possible to get an idea of (and compare) the main types of queries entered by users both within and across sites.

Overall, popular genres of queries included proper names, general subjects, locations, and topics pertaining to time (e.g. a year or a historical period) (see Table 2). However, the characteristics of queries appeared to be influenced by the subject domain. The most notable differences were between the historical domain and that of fine arts. While there is naturally some overlap between the two domains, queries to the historical domain (as represented by the St Andrew's collection and the "history" subsection of WIND) had more of an emphasis on place and time than their fine-arts correspondents (which were more heavily focused on named entities—i.e. the names of individuals or of artwork titles). A knowledge of characteristics of likely queries can assist the design of a translation system, as well as help to influence ways of letting users navigate a site.

Table 2: Categorization of top 100 queries from cultural heritage query logs

	Proper names	Subject	Place	Time
Tate	63	36	2	1
WIND	66	24	7	3
Cervantes	73	24	3	0
Alinari	n/a	-	-	-
St Andrews	10	25	64	0

Analysis of the categorization of 100 most common queries from various cultural heritage-related sites. Alinari's log files were not ranked in order of most popular, but a brief examination of the queries revealed that named entities, subjects, places and times were all represented, with a strong emphasis on place and subject.

Finally, scenarios developed to cover the three user groups being targeted and have been used to communicate the MultiMatch goals and proposed functionality with respect to these user groups. They also help to clarify within the project the most desirable system features and design options. For example, an extract from the original scenarios can be seen below.

[11] http://orange.sims.berkeley.edu/cgi-bin/flamenco.cgi/famuseum/Flamenco

> Leonardo is a content broker at BigSearchImages Ltd. He helps clients to find special content in BigSearchImages' image base. In the past Leonardo has used a personal set of archives (Getty Hulton, Alinari, Corbis, Bridgeman, etc.) to find the requested contents. He had to search the same contents in many different vertical repositories; this process was very time-consuming. He now uses MultiMatch to collect contents from different authoritative sources from a single search interface. MultiMatch also stores the queries done by Leonardo in the past.
>
> One client, who is writing a book, has just asked for some historical photos of every-day work in early industry, which she needs as illustration for her book.
> Leonardo can retrieve pictures from the MultiMatch content base using a thesaurus. It contains some keywords that match with the client's request. Leonardo can browse the resulting subset of images or further limit it by adding query expressions.

Some of these scenarios were chosen for validation with cultural heritage experts at a later stage in order to affirm their authenticity. These scenarios will be used as examples of use in future testing and evaluation of the system.

IX. Conclusions

This article describes the approach used to gather initial user requirements for the MultiMatch project. These requirements are valuable in facilitating the design and implementation of the MultiMatch system. Identifying the appropriate user groups and individuals is a key step when defining system requirements; the cultural heritage partners have worked at identifying the most relevant user groups likely to benefit from the services that will be offered by MultiMatch.

Different classes of users (from the educational, cultural tourism and cultural heritage professional sectors) have been identified, together with an analysis of the activities they perform and the scenarios in which the MultiMatch search engine can be expected to operate. The analysis enables the identification of users' activities, their needs and the creation of scenarios to help visualize and discuss the provision of future services. The ultimate goal of the process is to guide system design by understanding and anticipating, where applicable, the needs of the communities interviewed.

References

[1] Agile (2003). http://www.agileallience.org/home (last accessed September 2006).

[2] Boehm, B.W. (1988). A Spiral Model of Software Development and Enhancement, IEEE Computer, pp. 61-72.

[3] Goto, K. and Cotler, E. (2005). Web ReDesign 2.0: Workflow that Works, New Riders Publishing.

[4] Hackos, J. and Redish, J. (1998). User and Task Analysis for Interface Design, Wiley Computer Publishing.

[5] Hoffer, J.A., Valacich, J.S. and George, J.F. (2005). Modern Systems Analysis and Design, 4th Edition, Prentice Hall.

[6] Ingwersen, P. and Järvelin, K. (2005). The turn: integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.

[7] Kleppe, A., Warmer, J. and Bast, W. (2003). MDA explained: the model driven architecture: practice and promise. Object Technology Series. Addison-Wesley.

[8] Krutchen, P. (2003). The Rational Unified Process: An Introduction, 3rd Edition, Addison-Wesley.

[9] Kotonya, G. and Sommerville, I. (1998). Requirements Engineering – Processes and Techniques, Wiley.

[10] Lethbridge, T.C. and Laganière, R. (2001). Object-Oriented Software Engineering: Practical Software Development using UML and Java, Second Edition, McGraw Hill.

[11] Maciaszek L.A (2005): Requirements Analysis and System Design, 2 ed. Addison Wesley.

[12] http://www.comp.mq.edu.au/books/rasd2ed/ReadersArea/LectureSlides/pdfpage.htm?Chapter%202=RASD2ed_Ch2.pdf

[13] Nielsen, J., (1993). Usability Engineering, Academic Press, Boston.

[14] Preece, J., Rogers, Y. and Sharp, H. (2002). Interaction Design: Beyond Human-Computer Interaction, New York, NY: John Wiley & Sons.

[15] Pressman, R.S. (2001). Software Engineering: A Practioner's Approach. McGraw-Hill.

[16] Rose, D.E. and Levinson, D. (2004). Understanding user goals in web search. In Proceedings WWW 2004, pages 13–19, New York, NY, USA, 2004. ACM Press.

[17] Sommerville, I., Sawyer, P. (1997). Viewpoints: Principles, problems and a practical approach to requirements engineering, Annals of Software Engineering. 3, pp.101-30.Appendix I