Project no. 033104

MultiMatch

Technology-enhanced Learning and Access to Cultural Heritage
Instrument: Specific Targeted Research Project
FP6-2005-IST-5

D7.1 Evaluation Methodology

Start Date of Project: 01 May 2006
Duration: 30 Months

Organisation Name of Lead Contractor for this Deliverable
ISTI-CNR

Version: Final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Table of Contents

Executive Summary

The aim of the MultiMatch evaluation activity will be to ensure that the individual components and the complete system prototypes are tested and that their development is assessed during the project life-cycle with respect to performance and usability (i.e. laboratory-based and user-centred evaluation). Component and system development will be monitored to ensure that the functional specifications defined in WP 1 are respected and performance evaluation will comprehend internal project assessment plus evaluation against external benchmarks. The usability testing will involve the MultiMatch user groups and a set of field trials will be organised.

This first deliverable of Workpackage 7 describes the approaches that will be used in order to evaluate the MultiMatch system. It focuses in particular on the definition of the methodology to be used for Prototype 1 both with respect to the laboratory-based and the user-centred evaluation; an update will be released at 24 months providing final details on the evaluation processes decided for Prototype 2. Deliverables 7.2 and 7.3, to be consigned at 18 and 29 months respectively, will describe the results of the evaluation of Prototypes 1 and 2 at component and overall system level. Deliverable 7.4, to be consigned at 30 months, will describe the results of the field trials with the operational system.

We adopt both laboratory-based and user-centred evaluation strategies in order to be able to assess as many aspects as possible of the system.

The former will allow us to evaluate component and overall system performance mainly from the perspective of retrieval accuracy. For the first prototype we will build a set of test collections to enable us to perform some project internal evaluation of the different system components, comparing results obtained using with various settings and configurations in order to be able to tune and improve performance. We will also duplicate some of the past experiments of CLEF - the Cross-Language Evaluation Forum[1] - so that we can compare performance against known benchmarks (as far as the availability of relevant data allows us). Our aim is to establish base-lines for the different components of the first prototype so that we can measure improvements achieved with the second prototype.

However, measures of retrieval accuracy do not always correlate well with user satisfaction and overall system performance. The user-centred evaluations will thus aim at measuring the level of user satisfaction with the system and will focus on assessing the various user interfaces and the overall system performance in terms of a number of factors: ease-of-use; speed of response; meeting of the requirements of diverse user groups, etc.. It should be noted that user-centred evaluation will have an exploratory nature in the context of the project, mainly for two reasons: (i) MultiMatch search services will be prototypes in the context of the project duration, working on limited amounts of data; and (ii) evaluation of the second prototype has to be carried on in less than three months. Consequently, usability testing, interviews with users and qualitative feedback will have a more primary role in the evaluation than comparative field trials using example user queries and monitoring user search behaviour.

[1] See http://www.clef-campaign.org/

1 Introduction

The aim of the MultiMatch evaluation activity will be to ensure that the individual components and the complete system prototypes are tested and that their development is assessed during the project life-cycle with respect to performance and usability (i.e. laboratory-based and user-centred evaluation). Component and system development will be monitored to ensure that the functional specifications defined in WP 1 are respected and performance evaluation will comprehend internal project assessment plus evaluation against external benchmarks. The usability testing will involve the MultiMatch user groups and a set of field trials will be organised.

This first deliverable of Workpackage 7 describes the methodologies that will be used in order to evaluate the MultiMatch system. Deliverables 7.2 and 7.3, to be consigned at 18 and 29 months respectively, will describe the results of the evaluation of Prototypes 1 and 2 at component and overall system level. These results will be used to further tune and improve the performance of the system. Deliverable 7.4 at 30 months will describe the field trials performed on the final version of Prototype 2.

We intend to adopt both laboratory-based and user-centred evaluation in order to be able to assess as many aspects as possible of the system. The former, as described in Section 2, allows us to evaluate component and overall system performance mainly from the perspective of retrieval accuracy. For the first prototype we will build a set of test collections to enable us to perform some project internal evaluation of the different system components, comparing results obtained using various settings and configurations in order to be able to tune and improve performance. We will also duplicate some of the past experiments of CLEF - the Cross-Language Evaluation Forum[2] - so that we can compare performance against known benchmarks (as far as the availability of relevant data allows us). Our aim is to establish base-lines for the different components of the first prototype so that we can measure improvements achieved with the second prototype. However, measures of system effectiveness on the basis of retrieval accuracy do not always correlate well with user satisfaction and overall system performance. We will thus also conduct a series of user-centred evaluations, as described in Section 3, in order to assess the user response to the system. We will focus on evaluating both the various user interfaces and the overall system performance in terms of a number of factors: ease-of-use; speed of response; meeting of the requirements of diverse user groups, etc..

A correctly conducted laboratory-style evaluation should provide measures which are as objective as possible and can be compared against the current state of the art for similar systems. Because of the many different external factors involved in user-centred evaluation, despite the fact that we will follow established methodology, we expect far more subjective results, indications rather than ground truth.

The deliverable is divided into two main sections: Laboratory-based evaluation; User-centred evaluation including Field trails. Each section will describe briefly the methodology we intend to adopt for both Prototype 1 and Prototype 2. Although at the time of writing, we have clear ideas with respect to the evaluation of the components for the 1st Prototype, during the course of the project it is expected that revisions will be made with respect to evaluation for the 2nd prototype. For this reason, only the evaluations to be conducted for Prototype 1 will be described in detail. An update to this document will be produced at 24 months, defining the tasks and experiments for the evaluation of Prototype 2.

2 Laboratory-based Evaluation

The laboratory-based evaluation methodology that will be adopted in MultiMatch is based on the experience acquired by the academic partners in the coordination of the Cross-Language

[2] See http://www.clef-campaign.org/

Evaluation Forum. For its core activities, CLEF adopts a corpus-based, automatic scoring method, based on ideas first introduced in the Cranfield experiments in the late 1960s [Cleverdon, 1997]. This methodology is widely used and accepted in the information retrieval (IR) community and has since been used in major IR evaluation campaigns around the world such as TREC[3] (US), CLEF (Europe) and NTCIR[4] (Asia). Its properties have been thoroughly investigated and are well understood. CLEF has adapted it for application in a multilingual and cross-language retrieval context [Braschler & Peters, 2004]. "End-users" are not directly involved when following the "Cranfield" paradigm, the evaluation is strictly limited to examining certain parameters of the system performance which lend themselves most easily to an "objective" assessment and, in particular, the assessment of performance focuses on how accurate the system is in finding documents of interest according to a given specification of a user's information need, and ranking them in order of relevance.

Laboratory-based evaluation is popular not only because of the well-known costs and complexities incurred by conducting user-based evaluations but also because of the inherent difficulties in interpreting the results obtained with users. While it is important that the end users are involved in the system testing and we will certainly do this in the evaluation of the MultiMatch prototypes (see Section 3 below), abstracting the evaluation process can help to identify and control some of the parameters that affect retrieval performance.

2.1 Test Collections and Evaluation Metrics

The Cranfield experiments introduced the idea of creating test collections which could be used for comparative evaluation and results analysis. A test collection for IR system evaluation will consist of:

- a set of documents that are pertinent to the task in hand, i.e. as MultiMatch aims at building a multilingual multimodal search engine for cultural heritage, an ideal test collection will consist of documents in multiple language and diverse media relevant to this domain;
- a set of statements[5], simulating how users express their information needs, from which the system can derive queries that represent these needs, again pertinent to what aspect of the system is to be assessed;
- a set of relevance assessments, i.e. for each topic or query statement a list of the documents that are respond to the user need expressed.

The test collection used in the Cranfield experiments was small, consisting of approximately 1400 abstracts and 225 requests. However, evaluation using small collections often does not reflect the performance of systems in large-scale searching and does not demonstrate the ability of a system to operate in real-world IR environments. Following proposals first made by Karen Sparck Jones and Keith van Rijsbergen [1975], TREC has worked on establishing design criteria for the building of very large test collections (i.e. with hundreds of thousands of documents), more suitable to meet today's information needs if statistically significant results are to be obtained. CLEF has followed the directions proposed by TREC, described in [Harman, 2005] but has concentrated on building comparable test collections in multiple languages. For example, the main CLEF multilingual test collection consists of more than 3 million documents, comparable for topic and period, in 13 European languages, with associated topics and relevant judgments.

In order to conduct an evaluation exercise using the test collection, you need number of different participating systems in order to be able to compare performances and establish benchmarks. And

[3] Text REtrieval Conference series, http://trec.nist.gov/

[4] NTCIR (NII Test Collection for IR Systems) Project, http://research.nii.ac.jp/ntcir/

[5] Such statements in the TREC and CLEF vocabulary are known as "topics".

you also need metrics to measure the results. Popular measures usually adopted for exercises of this type are Recall and Precision. These are defined as follows:

Recall $\rho_r(q) := \frac{|D_r^{rel}(q)|}{|D^{rel}(q)|}$ and Precision $\pi_r(q) := \frac{|D_r^{rel}(q)|}{|D_r(q)|}$,

where $D_r(q) := \{d_1, ..., d_r\}$ is the answer set to query q containing the first r documents. The choice of a specific value for r is necessary because recall and precision are set-based measures, and evaluate the quality of an unordered set of retrieved documents. Choosing a low value for r implies that the user is interested in few, high-precision documents, whereas a high value for r means that the user conducts an exhaustive search. $D^{rel}(q)$ is the set of all relevant documents, and $D_r^{rel}(q) := D^{rel}(q) \cap D_r(q)$ is the set of relevant documents contained in the answer set [Schäuble, 1997]. When precision and recall are determined for every possible size of the answer set, a plot of the corresponding values results in a saw tooth curve. In the next step, typically a replacement curve is defined by assigning for every recall value $\rho \in [0,1]$ a precision value as follows:

$$\Pi_q(\rho) := \max\{\pi_r(q) | \rho_r(q) \geq \rho\}$$

Using this "interpolation step", we obtain a monotonically decreasing curve where each recall value corresponds to a unique precision value (see Figure 1). This "ceiling operation" can be interpreted as looking only at the theoretically optimal answer sets for which recall and precision cannot be improved simultaneously by inspecting further documents.

When evaluating a system with a set of queries (typically 50 in CLEF), an averaging step is introduced that produces the final recall/precision curve:

$$\Pi(\rho) := \frac{1}{|Q|} \sum_{q \in Q} \Pi_q(\rho)$$

where $|Q|$ denotes the number of queries.

However, often people prefer single value measures to a more complex performance indicator, such as a recall/precision curve. The advantage of such single value measures lies in easy comparison, their danger in too much abstraction: if relying exclusively on a single value, the ability to judge a system's effectiveness for different user preferences, such as exhaustive search or high-precision results, is lost.

The most popular single value measure for assessing the effectiveness of information retrieval systems is average precision. To calculate the average precision value, the precision after each relevant document found in the result list is determined as outlined above. The list of precision values that is obtained is then used to calculate an average. No interpolation is used to calculate the final average. When averaged over all the topics in a run, the measure is called the mean average precision (MAP).

[Buckley and Voorhees, 2005] state that average precision has a number of useful properties: contributions to the score are consistent with intuitive notions of what is important; a relevant document ranked highly contributes much more than a relevant document much further down the ranked list; the immediate contribution of each relevant document is known. Thus, MAP is a useful metric in failure analysis and system tuning. But they also indicate its weaknesses: in particular, that it is an overall system evaluation measure, not an application measure and there is no single user application that directly motivates MAP.

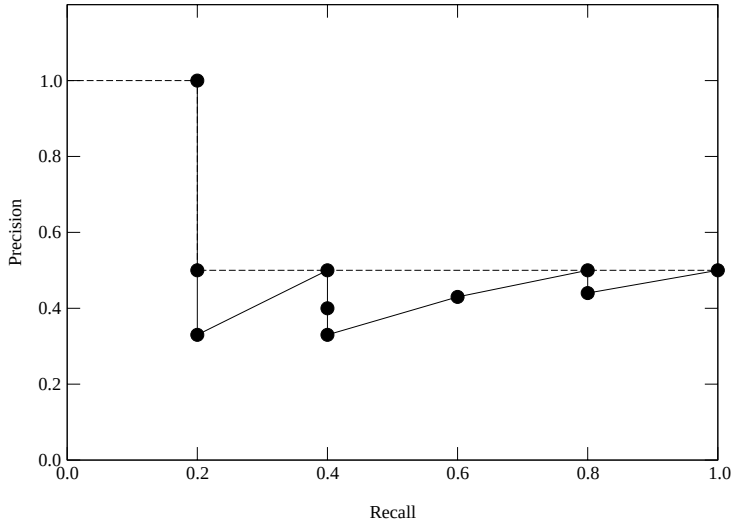

Fig. 1 Interpolation of recall/precision values

2.1.1 Test Collections and Metrics for Prototype 1

The building of test collections of the kind used in TREC or CLEF is a costly task in terms of effort whereas the resources of MultiMatch are of necessity limited. The most time-consuming task is generally relevance assessment as, even adopting the pooling methodology[6], this normally involves reading and assessing the relevance of hundreds of documents for each query.

For the evaluation of the first prototype we have thus decided on a two level evaluation:

- we will perform a project internal evaluation for separate system components (text, image, speech recognition) building our own test collections using MultiMatch content and using a known item search;
- we will test the same components externally using existing CLEF test collections and using an ad hoc style search.

Different evaluation metrics are employed for these two exercises: MRR and MAP.

A known item search is a particular IR task in which the system component is asked to find a single target document in a given document set. We will test for success within the first ten documents of the ranked list[7] and will use the mean reciprocal rank (MRR) to assess the results. An individual query will receive a score equal to the reciprocal of the rank at which the correct response was returned, or 0 if none of the responses contained a correct answer. The score for a submission is the mean of the individual queries reciprocal ranks. The reciprocal rank has several advantages as a scoring metric. It is closely related to the mean average precision measure used extensively in document retrieval. It is bounded between 0 and 1, inclusive, and averages well. A run is penalised for not retrieving any correct answer for a query but not unduly so [Voorhees & Garofolo, 2005]

In the ad hoc style search, a subset of relevant documents, rather than a single known item, are to be retrieved from a static test collection in response to a given query. MAP, as defined above, is the

[6] See [Voorhees, 2002] for a discussion.

[7] Success at 10 is chosen because it is well known that users generally expect to find information relevant to their query within the first ten documents and are rarely willing to scan further down the ranked list.

metric normally used to evaluate the results in this case and evaluation is performed using the trec_eval package made available by Chris Buckley[8]. External testing against CLEF benchmarks will give an indication of current component performance – however it has to be remembered that MultiMatch will be tuned for cultural heritage retrieval; the CLEF test collections are not CH specific. A detailed description of how we intend to perform this evaluation for the 1st prototype is given below (sections 2.2 and 2.3); an outline of our intentions for the 2nd prototype are given in 2.4 and 2.5.

2.1.2 Test Collections and Metrics for Prototype 2

We still have to decide exactly what test collections and evaluation metrics will be adopted in the laboratory-based evaluation of Prototype 2. In part, this will also depends on exactly which aspects of the system we intend to focus evaluation. The descriptions given in Sections 2.4 and 2.5 should thus be considered as purely indicative for now. The final decisions will be taken on the basis of the results of the evaluations of Prototype 1 and after further discussions between the partners. They will be reported in an update to this deliverable at 24 months.

2.2 1st Prototype – Component Evaluation

For prototype 1, laboratory-based evaluation will concentrate on component evaluation.

Mono- and Cross-language Text Retrieval

- Evaluation on known item search will measure performance in terms of the mean reciprocal rank measure; comparisons will be made between monolingual runs for the 4 core languages (Dutch, English, Italian and Spanish); cross-language runs over all pairs of languages against monolingual baselines.
  - The CH partners will be asked to prepare test collections for evaluation. DCU will provide example queries:
    - UA-BVMC – 50 queries on the Cervantes collection
    - BVMC / B&G / CNR will prepare 25 queries against Spanish / Dutch / Italian & English Wikipedia data.
    - All queries will be translated into all 4 languages by BVMC, B&G, CNR
- Monolingual evaluation using MAP against CLEF 2003 ad hoc benchmarks for the four languages; some cross-language testing using benchmarks for same year[9]

Image Retrieval

- Known item visual search and visual plus text search
  - Alinari will prepare 50 image queries and assess the results; USFD will provide example queries.
- We also intend to test the Image Retrieval component against the ImageCLEF St Andrews collection (historical photographic collection) benchmarks (2004 & 2005)

Speech Retrieval

- UvA will provide 25 speech queries (in English and Dutch) and will assess the results. DCU will provide example queries.
- We also intend to test the speech retrieval component against the CLEF MALACH speech collection benchmark (2006 & 2007).

[8] trec_eval 8.1, see http://trec.nist.gov/trec_eval/

[9] Cross-language evaluation in CLEF 2003 was only for certain language pairs:

2.3 1st Prototype – System Evaluation

The 1st prototype will be checked internally in order to see that the 1st prototype functional specifications have been respected; CNR and UNED will be responsible for this. Overall system evaluation will be done via demos and interviews with user groups – see 3.2 below.

2.4 2nd Prototype – Retrieval Components[10]

Mono- and Cross-language Text Retrieval

- The tests for the first prototype will be repeated expecting a considerable performance improvement
- We also intend to test the multilingual text retrieval component against the WebCLEF 2007 and 2008 benchmarks (the 2008 WebCLEF collection should consist of MultiMatch crawled content)
- The multilingual text retrieval components could also be tested against the benchmarks of CLEF ad hoc test collection for the relevant languages (although the MM text retrieval component has been tuned for domain-specific retrieval and thus the comparison is not balanced).

Image Retrieval

- The tests for the 1st prototype will be repeated and the mixed image + text retrieval components will be tested against ImageCLEF photo test collections (St Andrews 2004 and 2005).

Video Retrieval

- The video retrieval component could be tested against the benchmark established by the 2007 TRECVID competition – using only those queries that refer to CH content. The feasibility of this proposal will be assessed.

2.5 2nd Prototype – System Evaluation

It is our intention to organise a track in CLEF 2008 testing multilingual multimedia retrieval on cultural heritage content – using MultiMatch collections. The feasibility of this proposal will be discussed and a decision taken.

3 User-centred Evaluation

Evaluation of retrieval systems tends to focus on either the system or the user. Saracevic [1995] distinguishes six levels of evaluation for information systems (including IR systems):

- at the engineering level,
- at the input level,
- at the processing level,
- at the output level,
- at the use and user level and
- at the social level.

For many years IR evaluation has tended to focus on the first three levels, predominately through the use of standardized benchmarks (or test/reference collections) in a laboratory-style setting, as described in Section 2 above.

However, IR systems are increasing used in an interactive way within a social context, e.g. [Bates, 1989; Koenemann & Belkin, 1996; Rose & Levinson, 2004; Ingwersen & Järvelin, 2005; Rose, 2006] and this drives the need for user-centred evaluation to address performance at the latter three

[10] This description is to be considered as preliminary; an update will be released at 24 months

levels (output, user and use, and social). User-centred evaluation is important because it assesses the overall success of a retrieval system (as determined by end users of the systems) which takes into account other factors other than just system performance, e.g. the design of the user interface and system speed. Many researchers have argued that a system-orientated laboratory-based IR evaluation framework (like that typically undertaken in TREC) is not sufficient to test interactive IR and that alterative approaches must be employed, e.g. [Ingwersen & Järvelin, 2005:7]. A number of researchers have highlighted the need for user-centred evaluation in specific domains, e.g. [Dunlop, 2000] discusses an evaluation framework for evaluating interactive multimedia IR; [Hansen, 1998] discusses evaluation of IR user interfaces in web search; [Beaulieu et al., 1996] discuss evaluation of interactive IR systems within the TREC large-scale evaluation campaign and [Petrelli 2007] discusses evaluation within the context of interactive cross-language IR.

For this reason, the evaluation of MultiMatch search services will not be limited to system-oriented tests measuring the performance of MultiMatch retrieval components. A substantial part of the evaluation will be user-centred, oriented towards user satisfaction and efficiency and effectiveness of MM search facilities in user-oriented tasks.

Note, however, that user-centred evaluation will have an exploratory nature in the context of the project, mainly for two reasons: (i) MultiMatch search services will be prototypes in the context of the project duration, working on limited amounts of data; and (ii) Evaluation of the second prototype has to be carried on in less than three months. Consequently, usability testing, interviews with users and qualitative feedback will have a more primary role in the evaluation than comparative field trials using example user queries and monitoring user search behaviour.

3.1 Methodology

Evaluation of Interactive IR (IIR) systems typically forms part of an iterative design process (see, e.g. Hansen, 1998; Petrelli, 2007). Evaluation can be formative (run at any point during the design process) or summative (run at the end of a design project); undertaken within a realistic setting (e.g. a heuristic evaluation with experts in the field) or under controlled conditions (e.g. experiments in the laboratory with general users); operational (testing how a system performs in practice) or hypothesis-based (a research question formulated in advance and proved/disproved through evaluation); used to evaluate a system as a whole or individual components (e.g. a formative evaluation of sub-components of an IR system such as query formulation, browsing or results visualization) and use a range of data collection techniques (e.g. naturalistic or qualitative methods such as a content analysis of written data, or quantitative methods such as a statistical analysis of log data or questionnaire results).

User-centred evaluation in MultiMatch will be formative from the point of interface design and development (see Deliverable 6.1) and summative from the point of view of overall system evaluation (particularly of Prototype 2). We will undertake both a heuristic evaluation with experts in the field and a comparison with other search services in laboratory-based experiments under controlled conditions (in particular, with a fixed set of example search scenarios / tasks that users must solve given specific constraints, such as maximum search time). Evaluation will be mainly operational (testing how MultiMatch performs in practice) but hypothesis-based testing might also be attempted depending on the outcome of the MM research agenda.

An important part of evaluation is the definition of a suitable set of criteria and measures. Typically interaction is measured in terms of efficiency (e.g. search time), effectiveness (e.g. number of relevant documents collected) and user satisfaction, see e.g. [Su, 2003]. There is current research which suggested that evaluation of IR systems based on improvements in IR effectiveness measures only do not reflect on user's performance (e.g. accuracy and speed) [Al-Maskari et al., 2007; Turpin & Hersh, 2001]. The use of qualitative data in evaluation can be very helpful in explaining quantitative findings, see, e.g. [Petrelli, 2007] and gathering user feedback and comments on the IR system under evaluation (e.g. problems with the system, likes/dislikes etc.).

3.1.1 Evaluation in the User-centred MultiMatch Design Process

The design of the MultiMatch interface makes use of a user-centred design process (described more fully in Deliverable 6.1). Fig. 1 shows a typical iterative approach which includes gathering and analyzing user's needs/requirements, creating initial designs (e.g. low-fidelity mock-ups), evaluation and creating an interactive prototype. This has been done alongside WP1 as requirements are gathered, specified and refined. This design process involves using evidence from other sources to inform the interface design including materials such as related research projects and existing systems which exhibit similar functionality to MultiMatch, existing literature and the current functional specification (Deliverable 1.3).

The user-centred design involves consultation with representatives from user groups (e.g. educational, tourism and cultural heritage) to develop and evaluate a series of prototypes. Development is iterative and includes two main cycles for Prototypes 1 and 2. The development cycle includes the following stages: (1) needs assessment and task analysis; (2) preliminary design using low-fidelity prototypes; (3) design and development of interactive prototype; (4) heuristic evaluation and redesign; and (5) user evaluation. Steps 1-5 will be followed for the first prototype. The second prototype will follow a similar cycle, but starting from stage (3); that is, the needs assessment will not be carried out a second time.

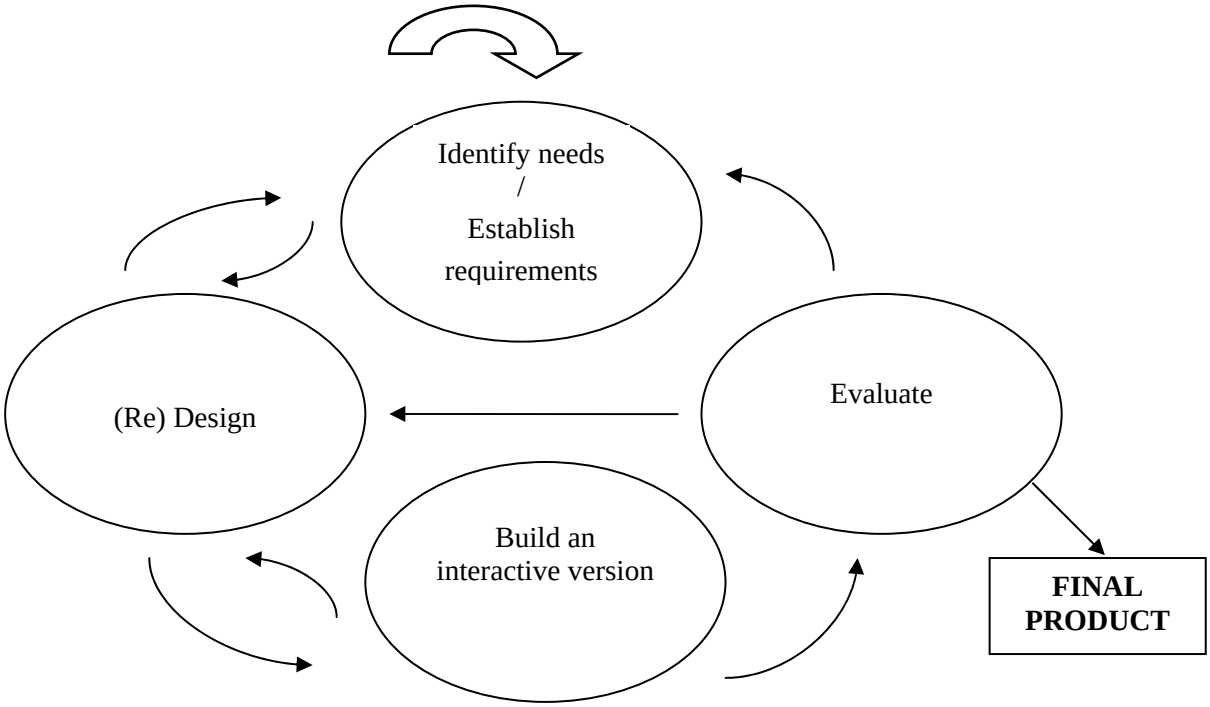

Fig. 2 An example iterative design process for user interface design [Preece et al., 2002]

What we describe in this report is step 5 of the process, the user evaluation. Initial evaluation with experts (heuristic evaluation) has already taken place as part of the design process (details in Deliverable 6.1). As advocated for the evaluation of IIR systems, see, e.g. [Hansen, 1998], we have already undertaken a study of end users and their typical work tasks/goals (task analysis), a crucial step in understanding (and designing) interaction for an IR system. Understanding the information seeking behaviour of end users (particularly between different user classes) helps to result in more effective interfaces which can support the user as he/she makes decisions in various information seeking tasks throughout the interaction process, see, e.g. [Robins, 2000].

The current evaluation is aimed at evaluating the first prototype (i.e. the interactive prototype) as a whole (testing sub-components are described within the design process). This corresponds to a summative evaluation of the MultiMatch system. While formal user testing will be carried out to evaluate the interactive prototype as a whole, there are some specific areas that will be the subject of particular focus for smaller scale experiments. For example, the degree and type of multilingual support desired by users with varying language skills and the type of functionality required for video visualization and playback. These and other experiments are being carried out both prior to and following the completion of the first prototype, and these will constitute a formative evaluation of the MultiMatch system.

The user test will be conducted once the final interactive prototype is finished. The results will prompt redesign and feed into the final first (or second) prototype designs. A controlled set of tasks will be set up in order to evaluate the system, most likely with general users. Users will be asked to complete these tasks (based on the processes described in the scenarios). By observing users during this process and obtaining their comments, it will be possible to identify those aspects of the system which are satisfactory and those which are unclear, confusing, or in need of changing.

3.1.2 Scenarios

Overall, the design is "task-driven" and developed to meet the needs of users carrying out prototypical tasks in the cultural heritage domain (called scenarios). In prior work on the interface design, a task analysis was carried out which attempts to understand what people currently do and why [Ingwersen and Järvelin, 2005; Hackos and Redish, 1998; Preece et al., 2002]. Part of the task analysis involves the creation (and refinement) of scenarios: an "informal narrative description" of human activities or tasks in a story or workflow [Preece et al., 2002: 222-234]. 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 scenarios can then be used during requirements analysis and to derive an understanding of the domain. In the design of the interface, the scenarios will assist with deciding functionality and evaluating the interface in later stages of the development.

Table 1 shows four scenarios resulting from discussions with cultural experts. These scenarios have been developed to provide a series of realistic tasks that users might typically perform with the MultiMatch system. These will be used to guide the user evaluations and provide context to search tasks given to end users (i.e. the search topics). This corresponds with the work of [Borland & Ingwersen, 1997] on simulated work tasks, the context in which user evaluations are performed.

[Shneiderman, 1998] describes a variety of potential information-seeking tasks: specific fact finding (known-item search,) extended fact finding, open-ended browsing, and exploration of availability. Previous task analysis with cultural heritage professionals revealed that specific fact-finding was the most common type of search activity; for this reason, the below scenarios are mainly focused on this type of behaviour. However, these may be supplemented with additional scenarios corresponding to other types of search behaviour (e.g. browsing.)

The use of scenarios in this evaluation framework aims to provide consistent tests enabling comparisons between the various user groups. The task descriptions in the following table are rather cursory descriptions; however, they will be expanded into narrative descriptions of personas for presentation to the evaluation participants. According to [Blomquist & Arvola, 2002: 197], the concept of a persona, as proposed by [Cooper, 1999], can be described as "an archetype of a user that is given a name and a face, and it is carefully described in terms of needs, goals and tasks."

Table 2. Scenarios prepared on the basis of interviews with cultural heritage experts

User type	Task	Media and languages involved
CH professional	Searching for video footage on Pier Paolo Pasolini, needs to gather background info on who he was	Text, Images, Video English, Dutch
CH professional	Looking for images of (non famous) people drinking coffee that capture a certain emotion	Images English, Italian
Academic	Preparing a presentation on Don Quixote and how it has influenced the arts	Text, Images, Video, Audio (?) English only
Cultural tourist/General user	Planning a visit to London, wants to know about museums to visit, what can see while there	Text, Images, Audio (podcasts) Spanish, some English

3.1.3 Example User Queries

Example user queries derived from log files provide a realistic set of topics to use for testing the system (both with user involvement and without, e.g. to test the various components.) A basic analysis of the log files of a range of cultural heritage website log files has already been carried out within the project by WP 6. This has provided a large set of realistic sample queries relating to the cultural heritage domain.

A simple analysis of the characteristics of the top 100 queries from each log file yielded the categorizations is shown in Table 3.

Sample queries chosen from these log files will come from some of the most frequently occurring terms submitted and will reflect the various common categories. Some examples include:

Named Entities: Leonardo, Pablo Picasso, Michelangelo, Don Quijote, Hamlet, Cervantes, The Last Supper, Francis Bacon, Plato, Caravaggio, van Gogh, Divine Comedy

Subjects: arte povera, still life, self portrait, surrealism, cinema, Gothic architecture, Etruscan tombs, surrealism, revolution, pop art

Places: Italy, Florence, Rome, La Scala, Arena of Verona, Colosseum, Glasgow, Louvre, Vatican, Milan, Loch Lomond

Time: Middle Ages, 1980s, World War I, Renaissance, 1960s

Table 3. Analysis of top 100 queries from CH websites

	Named Entities	Subject	Place	Time
Tate Online	63	36	2	1
WIND*	66	24	7	3
Cervantes	73	24	3	0
Alinari	33	29	39	1
St Andrews**	10	25	64	0

*This is a general purpose website; only CH-type queries were considered for this analysis

** These log files are from St Andrews University Library – a nationally significant collection of Scottish photography

3.1.4 Usability Testing

In order to evaluate the general usability of the MultiMatch site, a heuristic usability inspection will be conducted by an expert evaluator. The adherence of the system to basic usability guidelines (e.g., those proposed by [Nielsen, 1994] and/or [Shneiderman & Plaisant, 2004]) will be assessed. In addition, analytic studies will check the functionality of both prototypes against their functional specifications.

3.1.5 Naturalistic log analysis

The MultiMatch search service will be provided to Libero Portal users; this will provide logs of real search sessions, where the user population and the search needs are not laboratory-controlled conditions.

User activity with the system in field will be recorded with respect to issued queries and clicks on items of the result-set. End-user queries will be logged by the query engine, and a separate process will be developed to log end-user clicks.

Queries and clicks log-files will be managed in an ETL (*Extract, Transform, Load*) pipeline and data in the warehouse will be then analysed. Data will be mined to extract:

- The comparative usage of the various MM search services (for instance, use of innovative services versus standard ones).
- Query refinement induced by the MM advanced search facilities.
- the number of queries which returned an empty result-set;
- the distribution of the number of items in the result set;
- the number of "wasted" pages (i.e. the number of result-set pages the end-user did not click on);
- the distribution of the ordinal position of the clicked item(s) in the result-set.

3.1.6 Comparison with Other Search Services via Field Trials

There are two ways in which MultiMatch can be evaluated: either in comparison with another system, or as a standalone entity. Some previous work carried out as part of the formative evaluation has involved the use of other systems (e.g. Google Translate.)

The overall methodology used to compare MM prototypes with other search services will be a within-subjects design. Data collection and analysis methods will be based on those presented by [Hansen, 1998]. Data collection methods will include observation and measurement of objective criteria, along with questionnaires, in order to collect a combination of quantitative and qualitative data.

Users will be presented with the various scenario-based tasks to complete. In order to minimize order effects, tasks will be randomly counterbalanced using a Latin square matrix. Relevant quantitative measures related to these tasks (e.g. time taken) will be collected. However, as suggested by [Su, 2003], further subjective measures relating to user satisfaction (with difficulty, speed, coverage of results, etc.) will also be gathered through the completion of a short questionnaire following each task. Other input about user demographics and general feedback on MultiMatch will be solicited at some stage in the overall evaluation process.

The above procedure will initially be focused primarily on general users. However, some input from expert users has already been gathered earlier in the design cycle and will feed in to the redesign. The proposed evaluation methodology will also be run with some expert users (e.g. individuals at B&G and Alinari); results collected will also contribute to the evaluation.

3.2 Evaluation tasks and schedule

User-centred evaluation will focus on the MultiMatch user groups – described in D1.2.- consisting of educational users, tourists and cultural heritage professionals. Two main types of evaluation will be performed:

- Formative evaluation (intended to improve the final MM prototype, particularly in its interface aspects) will be ongoing throughout the project, under the responsibility of the Leaders of WP1 User Requirements (UNED) and WP6 User Interaction and Interface Design (USFD) with the collaboration of the cultural heritage institutions: Alinari, B&G and BVMC-UA. This type of evaluation will be reported in WP6 deliverables.
- Summative evaluation will be performed in the last three months of the project. The overall system performance will be evaluated by a series of field trials with groups of professional users, plus an analysis of the search activities of WIND Libero Portal users. This is the focus in the remainder of this section.

3.2.1 1st Prototype Evaluation

User-centred evaluation for the 1st prototype is conducted in 2 stages:

- the first stage regards evaluation of the interface and the proposed system functionality via discussions with small groups of users and using mock-ups (first results are already given in del 6.1)
- in the final stage – from 16 – 19 months (terminating with the MultiMatch workshop organised at EDUCA Berlin[11]) evaluation will consist of system demos and face-to-face interviews with users.

3.2.2 2nd Prototype Evaluation

The overall system performance will be evaluated in the last three months of project activity and will consist on a series of activities along the lines explained in the previous section:

- Analytic testing of compliance with the functional specifications. [Responsible: ISTI-CNR]
- Heuristic usability testing [Responsible: UNED]
- Heuristic evaluation via field trials (using example queries and search scenarios) and user interviews with groups of professional users as defined in D1.2. Feedback will be collected

[11] To be held on 28 November 2007, see http://project.alinari.it/diss-publish/mm_educaberlin.php

about usability, efficiency, effectiveness and adequacy to the different user scenarios. [Responsible: BandG. Participants: USFD and UNED, Alinari, BVMC].

- Log analysis of Libero Portal users. [Responsible: WIND]

  When the system becomes generally available, WIND will invite its web users via one or more of the following mechanisms:

  - direct invitation by email,
  - links placed in the "Libero Search" home page and specialized search
  - links placed in its focused portals (e.g."Canale Turismo") or the Libero Community homepage
  - query-based recommendations (certain CH-related queries might trigger a recommendation to launch the query in the MM search service).

  Users will be invited to use the service and complete a "web user" feedback survey (perhaps adapted from a subset of the expert survey). Two types of result will be given: results of the survey and the analysis of search logs (according to the lines introduced in the previous section).

  Note that, in order to optimize the outcome of the evaluation process, the system made available to the general public should be exactly the same as that evaluated by expert users, with the same functionality and the same contents. Content providers must agree in advance whether to replace IPR-protected contents by degraded or watermarked counterparts, or remove them altogether.

References

Al-Maskari, A., M. Sanderson, P.D. Clough (2007), The Relationship between IR Effectiveness Measures and Users' Satisfaction, In Proceedings of the ACM SIGIR2007 Conference (poster), Amsterdam, Netherlands, to appear 2007

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 407-424.

Beaulieu, M., Robertson, S., and Rasmussen, E. (1996). Evaluating interactive systems in TREC. J. Am. Soc. Inf. Sci. 47, 1 (Jan. 1996), 85-94.

Blomquist, A. and Arvola, M. (2002). Personas in action: ethnography in an interaction design team. In *Proc. NordiCHI2002* conference, pp. 197-200.

Borland, P., and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems, Journal of Documentation, 53(3) pp 225-250, June 1997.

Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, 7(1-2) pp 7-31.

Buckley, C & Voorhees, E.M. (2005). Retrieval System Evaluation. In Voorhees, E.M. and Harman, D.K. (Eds.). TREC. Experiment and Evaluation in Information Retrieval. MIT Press, 53-75.

Cleverdon, C. (1977) The Cranfield Tests on Index Language Devices. In: K. Sparck-Jones and P. Willett, eds. Readings in Information Retrieval, Morgan Kaufmann, 1997. pp 47-59.

Cooper, A. (1999). The inmates are running the asylum:Why high-tech products drive us crazy and how to restore the sanity. Indianapolis, Ind.: Sams.

Dunlop, M. (2000). Reflections on Mira: Interactive evaluation in information retrieval, Journal of the American Society for Information Science, 51(14), 1269-1274.

Forsyth, D.A. (2001). Benchmarks for Storage and Retrieval in Multimedia Databases. In Proceedings of SPIE International Society for Optical Engineering Vol. 4676, 240-247.

Goodrum, A.A. (2000). Image Information Retrieval: An Overview of Current Research. In Informing Science Vol. 3(2), 63-66.

Hackos, J., & Redish, J. (1998). User and task analysis for interface design. New York: Wiley Computer Publishing.

Hansen, P. (1998). Evaluation of IR User Interface – Implications for User Interface Design. Human IT. 2/1998. http://www.hb.se/bhs/ith/2-98/ph.htm

Harman, D.K. (2005) The TREC Test Collections. In Voorhees, E.M. and Harman, D.K. (Eds.). TREC. Experiment and Evaluation in Information Retrieval. MIT Press, 21-52.

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

Koenemann, J. & Belkin, N. (1996) A case for interaction: A study of interactive information retrieval behaviour and effectiveness. Proceedings of CHI96, 205-212

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), Usability Inspection Methods. John Wiley & Sons, New York, NY.

Petrelli, D. (2007). On the Role of User-Centred Evaluation in the Advancement of Interactive Information Retrieval. To appear in Information Processing & Management - special issue on "User-Centred Evaluation of IR Systems", Pia Borlund & Ian Ruthven eds.

Preece, J., Rogers, Y., and Sharp, H. (2002). Interaction design: Beyond human-computer interaction. New York: Wiley.

Robins, D. (2000). Interactive information retrieval: Context and basic notions. Informing Science, 3(2), 57-61.

Rose, D.E. (2006). Reconciling information-seeking behaviour with search user interfaces for the Web. JASIST 57(6): 797-799.

Rose, D. and Levinson, D. (2004). Understanding User Goals in Web Search. In Proceedings of WWW 2004, New York, USA. ACM.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

Schäuble, P.(1997): Content-Based Information Retrieval from Large Text and Audio Databases. Section 1.6 Evaluation Issues, Pages 22-29, Kluwer Academic Publishers, 1997.

Shneiderman, B. (1998). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, MA: Addison Wesley.

Shneiderman, B. & Plaisant, C., (2004). Designing the user interface: Strategies for effective Human-ComputerInteraction (4th ed.). Boston, MA: Addison Wesley.

Spark Jones, K. (Ed.). (1981). Information Retrieval Experiment. London: Butterworths.

Sparck Jones, K., and van Rijsbergen, K. (1975). Report on the need for and provision of an “ideal” information retrieval test collection. British Library Research and Development report 5266. Cambridge: Computer Laboratory, University of Cambridge.

Spark Jones, K. and Willett, P. (Eds.). (1997). Readings in Information Retrieval, San Francisco, CA: Morgan. Kaufmann Publishers, Inc.

Su, L. (2003). A comprehensive and systematic model of user evaluation of web search engines: I. Theory and background. JASIST, 54(13), 1175-1192.

Turpin, A. H. & Hersh, W. (2001). "Why batch and user evaluations do not give the same results" In: Proc ACM SIGIR, 225 - 231 New Orleans, Louisiana, United States

Voorhees, E.: The Philosophy of Information Retrieval Evaluation. (2002). In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval

Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers, 2002, Pages 355-370

Voorhees, E.M. and Garofolo J.S. (2005) Retrieving Noisy text. In Voorhees, E.M. and Harman, D.K. (Eds.). TREC. Experiment and Evaluation in Information Retrieval. MIT Press, 183-197.