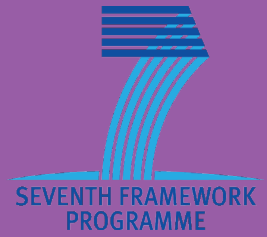
PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 2.1
Initial specification of the evaluation tasks

Version 1.0, 28th February 2011



Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

PROMISE
Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

SEVENTH FRAMEWORK PROGRAMME

Document Information

Deliverable number:	2.1
Deliverable title:	Initial specification of the evaluation tasks
Delivery date:	28/02/2011
Lead contractor for this deliverable	SICS
Author(s):	Jussi Karlgren, Gunnar Eriksson, Madlen Frieseke, Maria Gäde, Preben Hansen, Anni Järvelin, Mihai Lupu, Henning Müller, Vivian Petras, Juliane Stiller
Participant(s):	SICS, HES-SO, UBER, IRF
Workpackage:	2
Workpackage title:	Stakeholders Involvement and Technology Transfer
Workpackage leader:	SICS
Dissemination Level:	PU – Public
Version:	1.0
Keywords:	evaluation, use case, benchmarking, validation

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.8	21/02/2011	Draft	SICS	Circulated to contributing partners and to internal reviewers; uploaded to project site
0.9	25/02/2011	Draft	SICS	Revised after more input from the use case owners
1.0	28/02/2011	Final	SICS	Final version after further partner comments.

Abstract

Evaluation of multimedia and multilingual information access systems needs to be performed from a usage oriented perspective. This document outlines use cases from the three use case domains of the PROMISE project and gives some initial pointers to how their respective characteristics can be extrapolated to determine and guide evaluation activities, both with respect to benchmarking and to validation of the usage hypotheses. The use cases will be developed further during the course of the evaluation activities and workshops projected to occur in coming CLEF conferences.

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Table of Contents

Document Information 3
Abstract 3
Table of Contents 4
Executive Summary 5
1 Information Access Research is Based on Evaluation 6
2 Benchmarking and Validation 6
3 Use Cases Explained, Briefly 7
4 Variation Across Use Cases in Information Access 9
4.1 User factors 9
4.2 System factors 9
4.3 Source factors 9
4.4 Session factors 9
factor 10
typical values 10
relation to evaluation 10
User 10
5 Promise User Categories 11
5.1 Internal users: ”Researchers” 11
5.2 External users: “Consumers” 12
5.3 Beneficiaries: “Stakeholders” 12
6 Promise Use Cases 12
6.1 Use case domain “Search for innovation” 13
6.2 Evaluation criteria: recall, diversity 15
6.3 Use case domain “Unlocking culture” 17
7 Ramifications for Evaluation Activities 19
8 Industrial Relevance 20
9 List of Terms 20
References 21

Executive Summary

Information access research and development, and information retrieval especially, whatever the media type under consideration, is based on quantitative and systematic evaluation as the main vehicle of research. Evaluation of information systems typically proceeds by benchmarking system performance with respect to some gold standard but must - to be practically useful - also include a step of validating starting points and assumptions of effectiveness and usefulness through field studies or other forms of contact with users and usage situations.

Numerous factors related to usage, context and situation will influence the usefulness of a system for users, many of which are likely to influence the evaluation and the value of standard benchmarking tests. A non-exhaustive selection of factors is discussed in this deliverable. Some of these need to be studied further in realistic usage contexts.

In this deliverable, some initial formulations of use cases from the three PROMISE use case domains - : “Search for innovation” in the general area of patent retrieval, “Clinical decision support” in the general area of medical image retrieval, and “Unlocking culture” on accessing cultural heritage information sources - are given with the objective of giving a view of how a system might be useful for consumers outside the research laboratories: we here model processes and machinery which will be useful for deployed systems.

The ambition of PROMISE is to provide evaluation frameworks for each such use case and that evaluation labs in future CLEF cycles will be use case based, in order to provide for reusability and sustainability of results and cross-domain evaluation models.

In this deliverable first steps towards such formulations have been taken. In further cycles of formulation, more elaborate and formal formulations of use cases and the actions between user and system will be given.

1 Information Access Research is Based on Evaluation

Information access research and development, and information retrieval especially, whatever the media type under consideration, is based on quantitative and systematic evaluation as the main vehicle of research. Most typically, the evaluation follows the Cranfield model (Cleverdon and Keen, 1966) which is a benchmarking practice. A test set of pre-assessed target documents is used as a benchmark or gold standard for some collection, under the assumptions that

- an information need can be formulated satisfactorily and appropriately by the user;
- documents can be assessed as being relevant or not (or more or less relevant) for some given information need;
- the relevance of a document with respect to that information need is independent of other documents in the collection, based solely on the qualities of that document.

A system can then be evaluated after how well it delivers results in conformance with the benchmark. This abstracts evaluation away from variation of factors such as task, situation, context, user preferences or characteristics, interaction design, network latency and other such system-external qualities, systematically and intentionally ignoring factors relating to human behaviour and human interaction with information systems. This is good practice and has served the field well over a period of time within which information retrieval has positioned itself as one of the most important application areas of information technology and computer science.

2 Benchmarking and Validation

Benchmarking is only one part of evaluation. The original metaphor of benchmarking is useful to understand the point: bolting a piece of machinery to a workshop bench and running it with various inputs. Validating the starting points is as important: investigating if tools and technologies (and the design principles behind them) actually work for the tasks they are envisioned to address — if the machinery delivers performance when it is moved from the workshop into the production environment it is designed for. In the information access field, this means testing a system through user studies.

Performing valid user studies well is a craft in itself. To be of any impact, the user study must incorporate the crucial factors that can be expected to influence usefulness of the system under study. Laboratory user studies often implement an end-to-end system — including an interesting and newly developed piece of machinery — and have a number of test subjects use the system for a brief while in a laboratory environment with more or less realistic tasks assigned to them. This sort of study may be useful to evaluate the ergonomics of some specific interface widget, but they certainly are very unlikely to provide purchase to establish the validity of the starting points of a design for a task, whether the task is newly identified or a traditionally known one: the confounding factors in a

subsequent production environment majorise the variables studied in a typical laboratory test setup. As an alternative to laboratory tests, some variables and some hypotheses must be studied in the field or in field-like conditions or by observing practice in the field as the tasks in question are performed today.

Validating systems through studies is a challenging task. Doing this needs the hypotheses behind the system design to be explicit and operationalisable as study objectives. The system designers are often not the best professionals to execute valid user studies but the user study professionals are conversely not usually aware of what underlying hypotheses have informed the design of the system they see -- and then they cannot formulate the most appropriate evaluation study.

This is especially true when information access technology moves from its current prototypical domain of topical text retrieval, following the advent of multimedia as a large information carrier. Multimedia is different, used differently, by different users, and for different reasons than text. Benchmarking must change to capture the most important criteria for success for multimedia information access systems, using e.g. appeal, confidence, and satisfaction rather than completeness and precision as target notions — but when benchmarking changes, we risk losing the generality and sustainability of current information access evaluation results. If every multimedia search project comes up with its own user study, the results will be very difficult to compare across systems.

It is for this purpose we here propose a use case based approach to evaluation. If the system with its various technological features is evaluated with respect to the designers' hypotheses of how the system should be utilised, the evaluation can proceed to validate or disprove those hypotheses without being distracted by confounding factors.

Use cases may be put together on very various levels of ambition, competence, and insight, but once formulated, interaction specialists can debate and test the validity of the use case; information system specialists can set parameters for system benchmarking, based on crucial characteristics of the use case; and industrial and commercial stakeholders can use a validated use case to build and design their systems with benchmarked system components for their purposes, once they find it conforms to their business case.

3 Use Cases Explained, Briefly

Use cases are a relatively informal description of system behaviour and usage, which is designed to show how a system provides some value for the user when it is used. (Jacobson 1987, Jacobson et al 1992, Cockburn 2002, Övergaard and Palmqvist 2004) Use cases are used in systems design to identify, clarify, and organize system requirements. They are more or less informal, technologically neutral descriptions of typical ways in which the intended users will use the system. They define the actors – stakeholders, consumers, other systems who act outside the system being described - and the flow of actions to

accomplish a goal or a task of the primary actor. In other words, use cases treat the system as black box: describe what the system must accomplish, without saying how the system is to do it or occasionally as a "gray box" with some non-technical description of obviously crucial system components.

A use case is intended to capture all the ways a system is used by its environment, to describe all the services it offers and all the behaviour of the system and the actors engage in, for some specific purpose. The use case is a tool for developing a system, and user actions as formalised in the use case — most often using UML, the Unified Modeling Language — are mapped onto system components and system development objects for the purposes of system development and evaluation.

Scenarios, which often are the inspiration for use cases, are not use cases but instances of them: often several scenarios are necessary to track the various paths through a given use case for a system.

A scenario describes the actions of a user during the course of an interaction. For instance, one scenario based on the use case "search for image of friend" in a description of an image search engine could be a story of Pyramus entering his friend Thisbe's name in the query field of the interface to find an oil painting by J W Waterhouse[1] of her.

[1] Image of Thisbe from http://www.jwwaterhouse.com.

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

4 Variation Across Use Cases in Information Access

During the course of the European CHORUS coordination action a number of Europe-wide and national research projects were polled for their respective view of future usage of the technology solutions they proposed. The responses were aggregated and collated in project deliverables with the purpose of improving project-to-project cooperation. (CHORUS 2007, 2008; King and Kompatsiaris 2008) Table 1 gives some of the most salient features of use cases found in the CHORUS survey.

4.1 User factors

Factors directly related to the user or users have obvious implications for the evaluation. Two examples here will suffice: firstly, recent studies in collaborative IR [9] show how collectives of collaborating users break some of the patterns of single-user interaction with an information system. Evaluation of results cannot necessarily be done using metrics for individual retrieval. Secondly, the expertise of a user in domain or in the search system has immediate effect on evaluation: if a system is intended for professional users, a lab study with one session will not evaluate the long-term suitability of the solution in a professional setting and a probe study may be more appropriate and the system behaviour must be measured over a longer time depth or over a session rather than over a single search request.

4.2 System factors

Factors related to the technology used for interaction with the system, both as regards interaction device as well as the infrastructure for information transport will influence the presentation, the flow and the optimal configuration of information delivery. For instance, in the event of an information retrieval system, the size of screen and the convenience of input from the user - e.g. keyboard or voice input - will influence what result sets are likely to be most acceptable to the users.

4.3 Source factors

In interaction with information sources different from the prototypical text document collection a number of central factors of user satisfaction and thus evaluation change. If the interaction is with inherently streamed data, a database of retrospective material will become unrealistic and the current requirement of benchmarking to be reproducible on the same data set counterproductive. A more suitable requirement could conceivably be to require the benchmark results to be stable and predictable given some sampling procedure on the data stream. Additionally, if the source repositories are commercial and require users to pay for access to each item, the evaluation must incorporate a cost factor.

4.4 Session factors

The most obvious factors which motivate a separation between validation and benchmarking are here grouped under the heading session factors - factors which influence the interaction design of an information access session. These are factors such as dialogue

initative: is the system pushing information on to a possibly less committed user; are heavily engaged users putting great effort into finding the perfect fit of information to their need? Is the user attempting to retrieve a known item (in which case precision at one is the targeted evaluation criterion) or is the user browsing a collection to gain overview or to establish social relevance? Is the task the user is engaged in a professional task with external pressure for the user to perform well or an incidental and happenstance activity with of no lasting interest? Is there time pressure? Is the query formulation a simple or complex effort for users - are they likely to invest the effort given the retrieval performance of the system?

The factors given in the table are only a suggestion of the family of potentially crucial factors, the effect of which is likely in each case to majorise the currently minimal differences in mean average precision over a large set of recall points as an evaluation criterion.

factor	typical values	relation to evaluation
User		
social situation	single user; collaborative situation (synchronous/asynchronous; collocated/distributed; established group/adhoc group)	
domain expertise	novice vs expert	result ranking or selection
system usage	novice vs occasional vs expert	learning curve
System		
network	home / office / mobile	network latency; size of result
platform	personal computer / workstation / mobile device	size of result
Source		
media	text, audio, video, images, graphs, 3-D objects, maps, diagrams, data collections	gold standard set-up
business model	subscription, pay-per-view, no cost	cost calculation
repository	size, ownership, quality, provenance	browseability; quality and trust
permanence	collection vs stream	reproducibility
Session		
query	specification, example, set	formulation effort
initiative	push vs pull; lean-forward vs lean-backward	optimisation vs satisficing

factor	typical values	relation to evaluation
User		
context	none, implicit, user-specified, individual user model, stereotypical user model	fit over time to user model
goal	known-item search, overview, question answering, entertainment, socialisation, information refinement, monitoring	target notion: relevance / satisfaction / confidence
timeliness	real-time vs offline process	response time
persistence	single-shot, durational, repetitive	learning curve
result	single item, list (exhaustive or selection; ranked, ordered, organised), summary (report, overview, visualisation), answer (extraction, db fill), notification, browsing interface	recall-precision trade-off

Table 1: Non-exhaustive set of usage factors influencing evaluation methodology.

5 Promise User Categories

PROMISE is in the process of building a system for the research community in multimedia and multilingual information systems evaluation. The fact that the research infrastructure systems under construction largely are interactive and simultaneously are intended to provide tools, technologies and methodologies for interactive systems in use by end users entails a certain risk for terminological confusion. It is worth distinguishing between at least three user categories.

5.1 Internal users: ”Researchers”

Internal users of the research infrastructure are the primary users of the systems PROMISE will develop. WP3 and WP5 are concerned with developing tools for these users. We suggest these users are called researchers.

This includes developers, engineers, evaluators, annotators, assessors, track coordinators and many other roles that have to do with the research tasks we work with professionally. One of the stated aims of PROMISE is to increase participation in this sort of evaluation: we can envision more and different researchers in this role. An example of the latter could be future research activities based on the evaluation of campaign data accumulated and provided by the infrastructure system or companies using the data to compare systems or their components or generally following the research on the field.

User requirements (as formulated in WP 3 and WP 5) are necessary for the development process PROMISE engages in. These requirements are specific to the development of the

infrastructure for the activities PROMISE engages in - they will naturally be made available for anyone interested in similar development efforts, and while their details may be of less generality, the reasoning behind making design decisions can be expected be of lasting interest.

Researchers, unlike consumers (see below), are driven by the interest to study or evaluate a system - not by a certain information need.

5.2 External users: "Consumers"

External users of information access systems developed by the researchers are the users that are described in the use cases formulated in WP2 and are the targets for the evaluation metrics developed in WP4. We suggest these users are called consumers.

This includes professional users and searchers such as patent engineers; professional users without professional search training such as clinical practitioners and other professionals such as museum curators and archivists; laypeople and interested amateurs in the case of e.g. digital culture. These users are typically the clients of stakeholders defined below.

For the purposes of PROMISE we need to remember that test subjects are a proxy for consumers in laboratory exercises.

We want future tracks of CLEF to formulate use cases inspired by a vision of the needs of future users of practical fielded systems and sensitive to practicalities with respect of test subjects. These are what we mean by consumers.

5.3 Beneficiaries: "Stakeholders"

Third parties that have a valid interest and strong engagement in the development process PROMISE engages in are called stakeholders. Stakeholders are beneficiaries of the research infrastructure and especially of the research enabled by the infrastructure, and include information providers and producers, libraries, media companies, search engines and the like.

6 Promise Use Cases

PROMISE has as its starting point defined three use case domains: "Search for innovation" in the general area of patent retrieval, "Clinical decision support" in the general area of medical image retrieval, and "Unlocking culture" on accessing cultural heritage information sources. Each of these use case domains can accommodate numerous use cases, and PROMISE has chosen to elaborate one sample in each domain.

The ambition of PROMISE is to provide evaluation frameworks for each such use case and that evaluation labs in future CLEF cycles will be use case based, in order to provide for reusability and sustainability of results and cross-domain evaluation models.

In the following sections some very simple formulations of the first three PROMISE use cases are given with interaction sequences and annotations as to salient characteristics and questions for evaluation. Here the "user" in the use case should be understood in the "consumer" and "stakeholder" sense above: we are modelling processes and machinery which will be useful for deployed systems for external users. The needs of researchers and experiment leaders, the internal users of PROMISE, are taken care of in WP 3 and WP 5 of this projects.

The user actions are annotated with tentative consumer goals. This formulation of use cases is intended to show that if explicit hypotheses about consumer preferences are given, these hypotheses can guide evaluation both in choice of benchmarking metric and in making validation goal-directed.

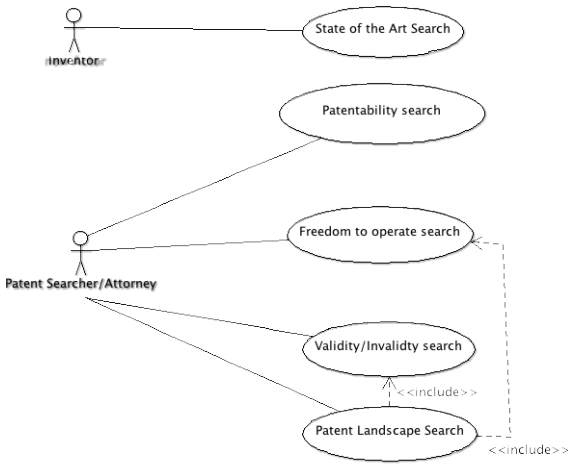
These use case formulations are not completely harmonised - the notation reflects the ongoing work in the respective use case domains. A near-future goal of the PROMISE project is to provide a framework for these and other future evaluation activities to specify the use case and attendant scenarios in a flexible semi-formal notation, inspired by standard UML notation but enhanced with the necessary fields and slots to guide evaluation activities.

6.1 Use case domain "Search for innovation"

Patents store large amounts of knowledge and give legal exclusive rights to the usage and implementation of inventions in our technology-driven world. To allow for an effective grant of new patents all past literature, especially patent publications, containing knowledge relevant to a certain domain need to be found. The "search for innovation" use case organizes evaluation campaigns where the consumers are inventors, patent searchers or patent attorneys performing a task related to their day-to-day activities, and described below. Regardless of the specific use-case describe below, a comprehensive patent search includes multi-lingual sources (i.e. a publication can be considered prior art regardless of the language it is published in), as well as multi-modal sources (e.g. for the chemical domain, images representing structures are often of paramount importance). In fact, patent search may be very different from one domain to another. For some, extensive domain specific indexes exist and multi-linguality is not a significant issue (e.g. chemistry). For others, images are paramount (e.g. engineering or patents which describe processes). In general, patent search is very domain specific.

The search for innovation use case organizes a lab in CLEF 2011 to compare state of the art multilingual retrieval techniques for finding all patents relevant to a particular topic. Part of the CLEF-IP 2011 lab is a pilot task organized in connection with ImageCLEF that aims at comparing visual retrieval techniques and the influence that visual information analysis can have for finding all relevant patents.

Several use cases are conceivable in the patent retrieval domain. The all share some characteristics that are bound to the domain and data under consideration: typically patent search is done by domain experts and professional information analysts in a professional setting using high-end office equipment, investing a non-trivial amount of effort on the formulation and specification of information need (although example-based queries might be conceivable) on the initiative of the user (although some monitoring-type usage situations are conceivable). The collective of users - in the case of collaboration - will be well established with professional credentials.



[State of the Art Search]
Objective: gain a comprehensive overview of a product or technology
When: before R&D investment has been done
On what: all available information sources
Date limit for searched publications: [-infinity, today]
Query: general request for information
Expected reply: large and comprehensive
Evaluation criteria: recall, diversity

[Patentability Search]
Objective: find all relevant prior art that may impact the likelihood of the patent being granted
When: before writing the patent application
On what: all available information sources
Date limit for searched publications: [-infinity, today]

Query: one or several claims
Expected reply: focused
Evaluation criteria: recall

[Freedom to Operate Search]
Objective: make sure that one does not infringe upon another's patent that is still in force
When: before a product is marketed/imported/manufactured
On what: granted patents in the target jurisdiction
Date limit for searched publications: [-25years,today]
Query: general request for information, but with technical details
Expected reply: focused
Evaluation criteria: recall

[Validity/Invalidity Search]
Objective: to determine if a patent already granted for an invention is valid
When: in case of litigation
On what: all available information sources
Date limit for searched publications: [-infinity, priority date of the granted patent+5years]
Query: one or several claims
Expected reply: focused
Evaluation criteria: recall

[Patent Landscape Search]
includes [Freedom to Operate Search] and [Validity/Invalidity Search]
Objective: to assess a company's patents - whether they are robust enough to exclude competitors and market the invention with the least probability of an infringement lawsuit
Query: a sequence of patents

6.2 Evaluation criteria: recall, diversity

Use case domain “Clinical decision support”: Medical image retrieval

Medicine is one of the most information-intensive fields and potentially affects all of us. Of all exams, imaging has created the largest amount of data available to physicians often with great benefit but also with a risk of data overload. Finding the right information and making it available to the right persons at the right moment is a challenge. Visual information retrieval is also still much less explored than textual information retrieval. The use case will thus focus on the visual information retrieval aspects and the inclusion of images. The medical literature currently constitutes an enormous knowledge base that includes visual as well as textual information.

Multilingual aspects equally play an important role in this domain as many people are more familiar with formulating information needs in their mother tongue even if they are understanding and speaking English, the language of most of the literature, well.

Several use cases are conceivable for Medical image retrieval - clinical usage, both urgent and non-urgent; research, both industrial and academic; students, in the medical field and in related fields; sensemaking for the general public. This use case domain concentrates on Clinical decision support -- on supporting a clinical practicioner performing a medical task.

The use case domain of Clinical decision support will organize a lab at CLEF 2011 to analyze the quality that current retrieval technologies deliver on retrieval from the medical literature in several languages and more particularly how visual information analysis can be integrated into the process in the best possible way. The evaluation will include practical demonstrations of retrieval systems that allow showing potential benefits and usability of such tools.

The use case for Clinical decision support involves a typically professional user, working alone or in a collaborative situation in an office or in a mobile situation across a large range of different data types. This sort of situation does normally not involve cost calculation on the part of the user: the data is either public or associated with the patient or patient group at hand; the repository is of high quality information.

Variation in this case is over the different types of query that can be posed. There is no inherent preference in the use case per se, as it is understood at this point, for any specific query type: specifications, examples, previously accessed sets of information, are all conceivable specifications of information need. Analogously, the result presentation can vary over single items, lists, summaries, database sets.

offline:

[data provision] data can come from several sources such as the local hard disk of a person, his PDA, from the electronic patient record based on access rights or from the Internet such as wikipedia or BioMed Central and other journals

[document translation and preparation] free text documents can be translated to be searchable in other languages and/or documents can be mapped to medical ontologies extracting symptoms, anatomic regions, modalities, pathologies, ...

online

[clinician]: has an information need as the situation of a patient is not 100% clear

[formulating a query] can be a precise query asking for textual information need, it can include structured data of the patient such as lab results and anamnesis and it can include one or several images

[pre-treatment of the query] data can be translated at this step or mapped to a medical ontology, images have their visual features extracted and structured data can be classified;

potentially this can include a definition of what the search goal is (textual fact, example image, similar cases)

[querying] separate queries can then be performed for the images and textual data

[results preparation] results of the separate queries can then be combined in various ways for calculating the ranking of the results depending on the included media and also the search goals; this step can indlue the translation of the results such as the abstract of an article or part of a patient records. Ontologies existing in several languages can be used for this

[results presentation] results will be presented to show the most important information, such as similar images, similar cases and/or results of a textual information need

[relevance feedback] based on the results a searcher can decide to reformulate the query and/or mark relevance feedback by selecting documents/images/cases as relevant or non-relevant to the initial query

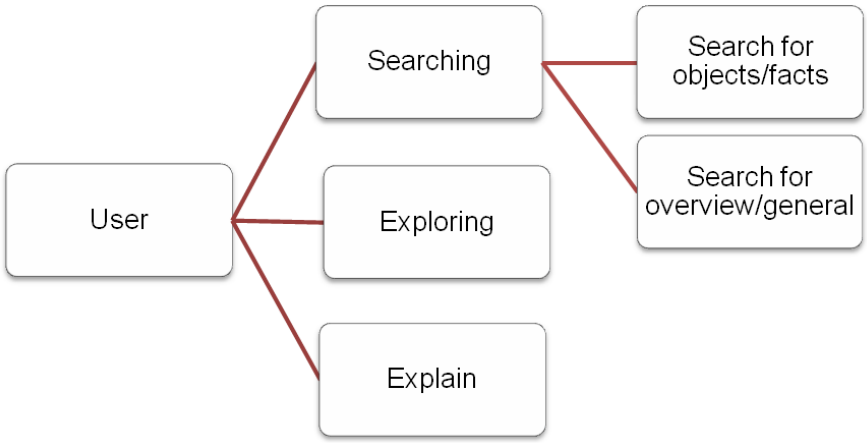
6.3 Use case domain "Unlocking culture"

The identities and distinctive features of most of societies are settled in their cultural heritage. Cultural heritage is strongly regional, particularly in Europe, comes in many different forms (books, paintings, sculptures, music, buildings) and is often language-dependent. The "Unlocking Culture" use case deals with effective information access to cultural heritage material held in large-scale digital libraries containing data from libraries, archives, museums, and audio-visual archives. Large quantities of cultural heritage objects have been digitized during the past few years in order to provide access to unique, rare or at-risk objects. However, access to these objects still poses several obstacles: the digital objects are provided through the metadata description efforts of the organizations and agencies curating the objects, usually in their national language and with specified technical vocabularies suited for their particular domains. Information systems for digital cultural heritage objects pose special problems related to the heterogeneous media types (texts, but more so images, audio or video files) and the uniqueness of the objects which makes their description difficult. Cultural heritage institutions have different approaches to managing information and serve diverse user communities, often with specialized needs. The scenario we are facing is to be able to satisfy user information needs by retrieving relevant "cultural assets" irrespective of the media type, location or language in which information objects are expressed.

Despite the fact that digital libraries are in a state of constant growth and much research is carried out in the field, much less is done to establish standard evaluation criteria and methods. For example, proceedings of the relevant conferences such as ECDL and JCDL contain no more than 5% of all papers related to the evaluation of cultural heritage information systems such as digital libraries.

The use case domain of "Unlocking culture" will be the central topic of the CLEF 2011 workshop CHiC 2011 Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage.

Users as well as user scenarios of Cultural Heritage systems such as Digital Libraries are quite heterogeneous. The following description of use cases start from the user actions: search, exploring and explain and describe possible patterns. There are no clear boundaries between the use cases identified so far. According to different user goals the three main use cases can be further subdivided into specific scenarios with more details concerning user and system performance (for more information see Multimatch D1.2: 2006).



[Search - fact / object]
Objective: Find a specific fact or object
On what: all information sources (all media types)
Query: specific (text input)
Expected reply: focused, display full result
System feature: filtering, advanced search
User goals: find Bible in English; find 'Mona Lisa',
Evaluation criteria: precision

[Search – Overview / General]
Objective: find all matching records
On what: entire collection of a site
Query: Broad / general request for information (text input)
Expected reply: broad across media types, display list of brief results
System feature: filtering, advanced search
User goals: find all works of an artist, find information about Renaissance, find pictures of Paris
Evaluation criteria: recall, diversity

[Browsing]
Objective: Thematic / subject access
On what: From general to specific
Query: Broad (Click / list presentation)
Output: matching browsing pattern
System feature: Facets, categories, tags, classification, controlled vocabulary, filtering
User goals: find artwork from specific provider; look at all pictures tagged with "flower"
Evaluation criteria: diversity, social relevance of displayed content

[Explain]
Objective: looking what the system offers / general interaction with system
On what: complete content of site including system pages and static and dynamic pages
Query: Click
Output: broad, depending on task
System feature: interactive interfaces
User goals: this use case includes all actions that cannot be assigned to the above mentioned; the user looks what is there; "Entertainment" is an important factor here
Evaluation criteria: user sense of satisfaction and completeness

As mentioned above these very general use cases can be divided and specified to scenarios describing a complete search or browsing process as produced for the Multimatch Project (D2.1: 2006 p. 26):

"After examining initial search results, Juan decides to improve his knowledge about the artwork "The Sunflowers" from Van Gogh a little bit more. He realizes that it was a previous search result showed by MultiMatch as a cultural object and decides to click on its link to see what happens. MultiMatch launches a new query based on metadata associated with "The sunflowers" cultural object and retrieves specific information about this topic. Juan realizes that MultiMatch has clearly separated and classified web pages according to general categories such as pages about the artwork, reviews of the artwork, news related with the artwork and noncategorized pages. He also can access a profile info box which describes the main features of the artwork. This is done by MultiMatch automatically."

7 Ramifications for Evaluation Activities

In a simplified example of how validation can be generalised from one scenario to another, consumers in a scenario for "Museum visit" might be established through a validation study to be found to prefer the system to switch from overview to in-depth lecture mode when they inspect an item in the museum collection more closely. Under the assumption they do,

evaluation of system components for that use case can proceed using a suitable benchmark method. Then if consumers in another scenario ”Library visit” also appreciate the system to go into in-depth mode if they select a specific literary work for further inspection - then system components that fit scenario ”Museum visit” can be used for scenario ”Library visit” as well, already having been benchmarked. However, a user study might confirm or disprove that hypothesis and instead show that library visitors prefer not to go into in-depth mode in which case the previously benchmarked system components must be re-evaluated or remain unproven for the task.

- For each use case and each system intending to contribute to it:
- Each user action in the interaction sequence needs to be motivated in terms of user goals which can be used to formulate evaluation target notions
- Each user action in the interaction sequence needs to be evaluated in terms of human factors
- Each system offering needs to carry suggestion of benchmarking and which parameters (may) be affected by user goals
- Each user goal needs to be validated
- Each system offering needs to be benchmarked with respect to user goal

In this deliverable first steps towards such formulations have been taken. In further cycles of formulation, sequence diagrams of actions between user and system should be formulated in terms which will help evaluation activities to be formulated.

8 Industrial Relevance

One of the success criteria for a successful evaluation of an information access solution is the ability to predict sub- sequent take-up of the solution in practice. The connection between benchmarking and take-up confounded by a large number of variables which may be difficult to model and the final quality of the complete system may hinge crucially on something completely different than the variables measured by benchmarking of its components. There is no reason to settle for anything but the best components, but if their effect cannot be measured in practice, it will be difficult to convince a commercial system designer to invest any effort in the improvements. Here, a validated use case with clear and explicit hypotheses of usage goals and linked to evaluation benchmarks will be a much more convincing argument than a benchmark alone.

9 List of Terms

Use case domain - the three given sample domains of PROMISE: Search for innovation, Clinical decision support, and Unlocking culture. Each serves as a basis for further development of use cases.

Use case - technologically neutral descriptions of how intended users will use the system, formulated in terms of actors and the flow of actions to accomplish a goal or a task of the primary actor. In PROMISE, use cases are intended to be enhanced with indications of how evaluation of use case oriented system solutions might proceed.

Scenario - an instance of a use case, with a descriptive narration to illustrate the interaction between system and user.

Sequence diagram - a sequence of actions in a use case in logical order.

Benchmarking - systematic, reproducible and quantitative comparison of system performance visavi some given standard measure, abstracting away from most user actions and other behavioural, contextual, and situational factors.

Validation - field- or empirically founded evaluation of starting points and basic hypotheses of user preferences and usage, as well as the effectiveness of technology and implementations as formulated in a use case.

References

Bennett 1971 — J. L. Bennett. “Interactive bibliographic search as a challenge to interface design”. In Donald E. Walker, editor, *Interactive bibliographic search: The User/Computer Interface*, pages 1–16. AFIPS, Montvale, New Jersey, 1971

Bennett 1972 — J. L. Bennett. The user interface in interactive systems. *ARIST*, 7:159–196, 1972.

CHORUS 2007 — Nozha Boujemaa, Ramón Compañó, Christoph Dosch, Joost Geurts, Yiannis Kompatsiaris, Jussi Karlgren, Paul King, Joachim Köhler, Jean-Yves Le Moine, Robert Ortgies, Jean-Charles Point, Boris Rotenberg, Åsa Rudström, Nicu Sebe. *State of the Art on Multimedia Search Engines.* CHORUS Project Consortium Deliverable D2.1. 2007.

CHORUS 2008 — Rolf Bardeli, Nozha Boujemaa, Ramón Compañó, Christoph Doch, Joost Geurts, Henri Gouraud, Alexis Joly, Jussi Karlgren, Paul King, Joachim Köhler, Yiannis Kompatsiaris, Jean-Yves Le Moine, Robert Ortgies, Jean-Charles Point, Boris Rotenberg, Åsa Rudström, Oliver Schreer, Nicu Sebe, Cees Snoek. *Identification of multi-disciplinary key issues for gap analysis toward EU multimedia search engines roadmap*. CHORUS Project Consortium Deliverable D2.2. 2008.

Cleverdon and Keen 1966 — C. W. Cleverdon and M. Keen. *Cranfield CERES: Aslib Cranfield research project - Factors determining the performance of indexing systems*. Technical report, 1966.

Cockburn 2002 — A. Cockburn. *Agile software development*. Addison-Wesley, 2002.

Jacobson 1987 — Ivar Jacobson. Object-oriented development in an industrial environment. *Proceedings of OOPSLA ’87: Sigplan Notices*, 22(12):183–191, 1987.

Jacobson et al 1992 — Ivar Jacobson, M. Christerson, P. Jonsson, and Gunnar Övergaard. *Object-Oriented Software Engineering: A Use Case Driven Approach.* Addison-Wesley, 1992.

King and Kompatsiaris 2008 — Paul King and Yiannis Kompatsiaris. *Towards a use case ontology for multimedia information retrieval.* CHORUS Consortium working paper. 2008.

Övergaard and Palmqvist 2004 — Gunnar Övergaard and Karin Palmqvist.*Usecases-Patterns and blueprints.* Addison-Wesley, 2004.