PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 2.2 - Revised Specification of Evaluation Tasks

Final version, February 2012

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:	2.2
Deliverable title:	Revised Specification of Evaluation Tasks
Delivery date:	29/02/2012
Lead contractor for this deliverable	UvA
Author(s):	Anni Järvelin, Gunnar Eriksson, Preben Hansen, Theodora Tsikrika, Alba Garcia Seco de Herrera, Mihai Lupu, Maria Gäde, Vivien Petras, Stefan Rietberger, Martin Braschler, Richard Berendsen
Participant(s):	SICS, HES-SO, TUW, UBER, ZHAW, UvA
Workpackage:	2
Workpackage title:	Stakeholders Involvement and Technology Transfer
Workpackage leader:	SICS
Dissemination Level:	PU – Public
Version:	Final version
Keywords:	Use cases, evaluation tasks

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

Abstract

Cranfield style evaluation has been dominant in evaluation of information access systems ever since the TREC evaluation campaigns, and successful as a platform for innovation. However, Cranfield style evaluations have limitations. Assumptions about typical end users, their tasks, goals, local environment and social context are often not made explicit. In PROMISE WP2 we develop a use case framework for explicitly describing use cases underlying evaluation tasks. The framework allows for describing very different use cases, broadening the scope of the traditional ad hoc search evaluation. We work out use cases from the three main use case domains studied in the PROMISE project (medical, search for innovation and cultural heritage), as well from the people search domain. For each use case, one or more evaluation tasks are discussed. The use case framework is not an evaluation framework. It allows a description of use cases, not evaluation tasks. This means that the framework can be used to inform evaluation experiments of any kind. The evaluation tasks, experiments and efforts described in this deliverable provide a case in point. One of the use cases, historical newspaper search, describes an associated evaluation experiment in which session based evaluation is performed. The way aspects of this evaluation setup are related to use case features showcases the potential of the use case framework to influence evaluation criteria such that evaluation reflects end user preferences better. Many other evaluation tasks in this deliverable are strongly rooted in the Cranfield tradition of test collection based evaluation. Therefore, we discuss some common aspects of test collections and relate them to relevant use case features. We provide some analysis of properties shared by our set of use cases and identify points for future work on new use cases. We plan to validate use cases in the sense that they should reflect usage by real end users of real services through interviewing these end users and service providers. In addition to evaluation tasks, we report on another evaluation effort, the black-box evaluation effort, which has been started in PROMISE independently from the development of the use case framework. It aims to evaluate entire information access applications from the perspective of the user in a systematic way. We show how it can be informed by and adapted to use cases described in the use case framework.

Table of Contents

Executive summary

Cranfield style evaluation has been dominant in evaluation of information access systems ever since the TREC evaluation campaigns, and successful as a platform for innovation. However, Cranfield style evaluations have limitations. Assumptions about typical end users, their tasks, goals, local environment and social context are often not made explicit. In PROMISE WP2 we develop a use case framework for explicitly describing use cases underlying evaluation tasks: the (desired) functionality of systems under scrutiny, typical end users, their tasks, goals, local environment and social context. It builds on and extends Section 4 of deliverable 2.1 (Karlgren et al, 2011), which contains a systematic discussion of variation across use cases in information access.

The use case framework we develop allows for describing very different use cases, broadening the scope of the traditional ad hoc search evaluation. A first version of the framework is presented in Section 2, while Section 3 contains worked out use cases from the three main use case domains studied in PROMISE (medical, search for innovation and cultural heritage), as well from the people search domain. For each use case, one or more evaluation tasks are discussed. The use cases are specified in more detail compared to deliverable 2.1 (Karlgren et al, 2011). They are examples of how the framework can be productively used in experimental design and reporting with a minimal threshold for adoption. The use case framework is intended to be of use also for future evaluation efforts outside the PROMISE project.

The use case framework is not an evaluation framework. It allows a description of use cases, not evaluation tasks. This means that the framework can be used to inform evaluation experiments of any kind. Any evaluation experiment can benefit from a systematic description of use cases involving the systems being evaluated. The evaluation tasks, experiments and efforts described in this deliverable provide a case in point. There is a task proposing session based evaluation, a task addressing diversity in the search result page, there are interactive evaluation tasks and there is an evaluation effort (the black-box evaluation effort) aiming to evaluate information access applications as a whole, rather than just the quality of the ranking produced by the search engine. All of these experiments can be informed by characteristics of use cases.

One of the use cases in Section 3.6, historical newspaper search, describes an associated evaluation task in which session based evaluation is performed (Keskustalo et al, 2009). Sessions are simulated from keywords which are obtained through a user study. The way aspects of this evaluation setup are related to use case features showcases the potential of the use case framework to influence evaluation criteria such that evaluation reflects end user preferences better.

Many other evaluation tasks in this deliverable are strongly rooted in the Cranfield tradition of test collection based evaluation. Most of the time, they target a specific subset of the desired functionality described in the use case associated with the evaluation task, typically a search engine. Therefore, in Section 2.5 we discuss some common aspects of test collections such as the collection, the topics selected, and the relevance assessments obtained and relate them to relevant use case features. The specification of the historical newspaper search evaluation task in Section 3.6 is an example where these links are often explicitly used to motivate choices in the evaluation setup. We believe test collection based evaluation tasks can be validated in this way to some extent, in the sense that we can investigate if evaluation outcomes will reflect user preferences.

A long term goal in applying this framework is to associate best practices for evaluation with use cases that share certain characteristics. The quality and quantity of such best practices will increase if many use cases are described in the framework, and the framework evolves further. In Section 3, we include only a limited amount of use cases. Nevertheless, we provide some analysis of properties shared by our set of use cases and identify points for future work in Section 4.

It is very important to validate use cases as formulated with the use case framework in the sense that they reflect usage by real end users, of real systems owned by real service providers (stakeholders). One approach we will take to work towards this goal is interviewing stakeholders and end users. We elaborate on this in Section 5.

In Section 6 we report on the black-box evaluation effort, which has been started in PROMISE independently from the development of the use case framework. It aims to evaluate entire information access applications from the perspective of the user in a systematic way. This work was started by Braschler et al. (2009), focusing on enterprise search. In the PROMISE project, the aim is to generalize it further on the one hand. On the other hand, the aim is to diversify it in order to accommodate very different use cases.

In the black-box evaluation setup, many system aspects, ranging from accuracy of metadata in the collection to entertainment of the user, are evaluated at the same time. Evaluation criteria from different categories (such as the quality of search results, the quality of the collection, and the quality of the user interface) can be gathered into lists of criteria that assessors can be instructed to use to rate the system under scrutiny. By applying different criteria and by weighing criteria differently this evaluation methodology can be adapted to different use cases described in the use case framework.

An important approach to evaluation discussed in deliverable 2.1 (Karlgren et al, 2011) is to conduct user studies. User studies are a very powerful way of controlling variables to isolate those variables that contribute to user satisfaction, task completion time or task accuracy. In contrast with benchmarking, user studies are very expensive in terms of labour. They are typically conducted with only a small amount of users. The fact that people are very

different and display unexpected behaviour becomes a challenge and limits repeatability of such experiments. We hope that a systematic description of use cases and the way they inform choices in the setup of evaluation experiments can help bridge the gap between user studies and benchmarking.

In Section 7 we discuss the current status of our work and identify next steps and short term goals for the next deliverable. After a few iterations, the use case sections and features seem to have stabilized. We can focus now on formulating hypotheses about user preferences starting from the use case features. For each feature, we can investigate how it relates to evaluation. Features may interact in this respect, complicating the matter. Once we succeed in making explicit which factors play a role in determining evaluation decisions, interaction specialists can start validating the underlying hypotheses made (e.g. through user studies), and information system specialists can adjust parameters for system benchmarking based on crucial characteristics of the use case. This is a long term goal of our work.

A next challenge for the short term in WP2 is to find a very condensed form for the specification of test collection based evaluation tasks (as opposed to use cases) that will enable organizers of these tasks to demonstrate how choices in evaluation setup are motivated by the underlying use cases. If we want to motivate organizers from outside the PROMISE project to do this, we will have to find a minimal set of use case features, aspects of evaluation tasks and relations between them that is still useful. Our aim for CLEF 2012 is to ask all lab owners for a one page specification of their evaluation task, with the underlying use case in mind.

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

1 Introduction

Cranfield style evaluation has been dominant in evaluation of information access systems ever since the TREC evaluation campaigns. Typically, search engines output a ranked list of search results in response to a free text query. Evaluation has focused on assigning a score to such a ranked list based upon the relevance of each returned document to the information need underlying the query. Information needs and relevance assessments are generally created by assessors, usually also with the help of search engines to locate potentially relevant documents. Creating topics and relevance assessments is labour intensive. However, once a test collection is in place, it can be used over and over again to compare a variety of systems. In addition, subtle differences in ranking quality can be detected, undetectable by end users. This provides a platform for accumulative, small improvements that ultimately should be perceivable by the end user. TREC campaigns have been a platform for innovation in this way from the beginning (Sanderson, 2010).

However, Cranfield style evaluations have limitations. The tasks evaluated are abstractions of real tasks. Assumptions about typical end users, their tasks, goals, local environment and social context are often not made explicit. But even if they are not, every test collection has an underlying user and task model. Every decision regarding topics, relevance assessments, metrics chosen reflects certain assumptions about a typical end user. For example, in ad hoc TREC campaigns, the end user is assumed to issue informational queries (Broder, 2002; Rose & Levinson, 2004), to have liberal relevance criteria (Sormunen 2002), and to find duplicates of already seen relevant information still relevant. In PROMISE WP2 we develop a use case framework for explicitly describing the use case associated with an evaluation task: the (desired) functionality of systems under scrutiny, typical end users, their tasks, goals, local environment and social context. It builds on and extends Section 4 of deliverable 2.1 (Karlgren et al, 2011), which contains a systematic discussion of variation across use cases in information access.

Ad hoc search evaluation tasks like the one described above can work well to establish the usefulness of systems with respect to some human activities if the activities in question fit this implicit use case. As it is uncertain that this specific use case would cover a large enough part of human information seeking activities to motivate evaluation based solely on it, it would make sense to look into other kinds of use cases too. The use case framework we develop allows for describing very different use cases, broadening the scope of the traditional ad hoc evaluation. A first version of the framework is presented in Section 2, while Section 3 contains the use cases. For each use case one or more evaluation tasks are discused The use cases are specified in more detail compared to deliverable 2.1 (Karlgren et al, 2011). The worked-through use case examples (Section 3) show how the framework can be productively used in experimental design and reporting with a minimal threshold for adoption. The use case framework is intended to be of use also for future evaluation efforts outside the PROMISE project.

The use case framework is not an evaluation framework. It allows a description of use cases, not evaluation tasks. This means that the framework can be used to inform evaluation experiments of any kind. Any evaluation experiment can benefit from a systematic description of use cases involving the systems being evaluated. The evaluation tasks, experiments and efforts described in this deliverable provide a case in point. There is a task proposing session based evaluation, a task addressing diversity in the search result page, there are interactive evaluation tasks and there is an evaluation effort (the black-box evaluation effort) aiming to evaluate information access applications as a whole, rather than just the quality of the ranking produced by the search engine. All of these experiments can be informed by characteristics of use cases. We address these four kinds of evaluation experiments in a bit more detail now. All of this work is relevant also for WP4 in PROMISE, which studies evaluation metrics and methodologies.

Recent years have seen an increased interest in evaluation user sessions as opposed to single queries as done in most evaluation tasks, leading to the first Session tracks to be organized at TREC in 2010 and 2011. One of the use cases in Section 3.6, historical newspaper search, describes an associated evaluation task in which session based evaluation is performed (Keskustalo et al, 2009). Sessions are simulated from keywords which are obtained through a user study. The use of simulation in the context of evaluation is an emerging research area. The way aspects of this evaluation task are related to use case features showcases the potential of the use case framework to influence evaluation criteria such that evaluation reflects end user preferences better.

Diversity in a result list makes relevance of items to queries dependent of the relevance of other retrieved items. The “Variability” evaluation task of the cultural heritage domain which will be organized at CLEF 2012 requires systems to return twelve items from the collection such that they form a good overview of relevant items from the collection. ‘Must-sees’ may be highlighted. The task is multilingual and multimodal. Search results should be diverse with respect to their content, but also with respect to their media type and content provider. Clearly, the task extends the typical ad-hoc search task in many ways.

In an interactive evaluation task, participating teams are assigned the same search task, and submit search results that have been obtained in any fashion, manual, or with the aid of various systems. PatOlympics is one effort that falls into this category, we reported on it in deliverable 4.1 (Berendsen et al, 2011). PatOlympics is an evaluation task designed for the ‘Prior art search’ use case in Section 3.2 below. In the medical use case domain running an interactive campaign is a long term goal. (Section 3.1)

A shortcoming of typical test collection based evaluation approaches is that only ranking quality is evaluated, while many other aspects of information access applications may contribute to user satisfaction. Moreover, Harman & Buckley (2009) signal that in ad-hoc search, absolute system scores have been flattening out in the last years of the TREC ad-

hoc campaigns. In such cases, investing in improving other aspects such as usability may offer more return on investment. In Section 6 we report on the black-box evaluation effort, which has been started in PROMISE independently from the development of the use case framework. It aims to evaluate entire information access applications from the perspective of the user in a systematic way. This work was started by Braschler et al. (2009), focusing on enterprise search. In the PROMISE project, the aim is to generalize it further on the one hand. On the other hand, the aim is to diversify it in order to accommodate very different use cases.

In the black-box evaluation setup, many system aspects, ranging from accuracy of metadata in the collection to entertainment of the user, are evaluated at the same time. Evaluation criteria from different categories (such as the quality of search results, the quality of the collection, and the quality of the user interface) can be gathered into lists of criteria that assessors can be instructed to use to rate the system under scrutiny. These criteria may be subdivided in such specific and simple tests that assessors can rate the systems rather objectively (not leaving much room for the assessors own perceptions and opinions). By applying different criteria and by weighing criteria differently this evaluation methodology can be adapted to different use cases described in the use case framework.

Many other evaluation tasks in this deliverable are strongly rooted in the Cranfield tradition of test collection based evaluation. Most of the time, they target a specific subset of the desired functionality described in the use case associated with the evaluation task, typically a search engine. Therefore, in Section 2.5 we discuss some common aspects of test collections such as the collection, the topics selected, and the relevance assessments obtained and relate them to relevant use case features. The specification of the historical newspaper search evaluation task in Section 3.6 is an example where these links are often explicitly used to motivate choices in the evaluation setup. We believe test collection based evaluation tasks can be validated in this way to some extent, in the sense that we can investigate if evaluation outcomes will reflect user preferences.

While benchmarking facilitates detecting significant system differences in average performance over a set of topics, even when these improvements are small, Cranfield evaluations fall short of explaining the large variance of individual systems over different topics (Harman & Buckley, 2009). In addition, since search engines are complex and different components may play a role in performance, treating and evaluating systems as black boxes may limit our understanding the interaction between queries and performance of systems. The DIRECT infrastructure developed in WP3 of the PROMISE project and the visual analytics tools developed in WP5 aim to facilitate extensive analysis of experimental results over many evaluation tasks, systems, queries and collections. The use case framework developed in WP2 can in the long term bring in properties of the underlying use cases of evaluation tasks to further inform this analysis.

Another long term goal in applying this framework is to associate best practices for evaluation with use cases that share certain characteristics. The quality and quantity of such

best practices will increase if many use cases are described in the framework, and the framework evolves further. In Section 3, we include only a limited amount of use cases. Nevertheless, we provide some analysis of properties shared by our set of use cases and identify points for future work in Section 4.

It is very important to validate use cases as formulated with the use case framework in the sense that they reflect usage by real end users, of real systems owned by real service providers (stakeholders). One approach we will take to work towards this goal is interviewing stakeholders and end users. We elaborate on this in Section 5.

An important approach to evaluation discussed in deliverable 2.1 (Karlgren et al, 2011) is to conduct user studies. User studies are a very powerful way of controlling variables to isolate those variables that contribute to user satisfaction, task completion time or task accuracy, For example, Turpin & Scholer (2006) show that ranking quality in terms of MAP does not necessarily correlate with task based measures such as task completion time or task accuracy. Smith & Kantor (2008) show that user adaptation may play a crucial role here, end users can obtain good results with bad systems by changing their interaction strategies. . User studies are very expensive in terms of labour, and are typically conducted with only a small amount of users. The fact that people are very different and display unexpected behaviour becomes a challenge and limits repeatability of such experiments. We hope that a systematic description of underlying use cases and the way they informed choices in the setup of evaluation experiments can help bridge the gap between user studies and benchmarking.

In Section 7 we discuss the current status of our work and identify next steps and short term goals for the next deliverable. After a few iterations, the use case sections and features seem to have stabilized. We can focus now on formulating hypotheses about user preferences starting from the use case features. For each feature, we can investigate how it relates to evaluation. Features may interact in this respect, complicating the matter. Once we succeed in making explicit which factors play a role in determining evaluation decisions, interaction specialists can start validating the underlying hypotheses made (e.g. through user studies), and information system specialists can adjust parameters for system benchmarking based on crucial characteristics of the use case.

2 A use case framework

Information access is inherently an interactive process, where individuals going about their business search for information to support their daily activities, whatever they may be. Recent decades have seen a growing understanding of work task requirements and effects on human information access. Also, there is a growing understanding that information access is broader in scope than problem solving in professional work task settings. At the same time, understanding on how to use this knowledge to derive and apply design criteria for information access systems has not advanced and information access remains a field

where a major part of the research efforts is directed towards algorithms for the computationally efficient representation and matching of sets of documents with terse queries. Information access evaluations are systematically abstracted away from factors related to users, their goals and system usage, and the user-related variables are either excluded or fixed in the experimental settings without much discussion or motivation. As a consequence, information access research results do not transfer well into real life contexts and their applicability as guidelines for practical system, product and service design is low. Thus, there is a clear need for an evaluation framework that can make explicit the hypotheses about user preferences, goals, expectations, and satisfaction that guide information access system evaluation.

In this section, an initial version of such a framework is presented. The framework is inspired by the use case methodology for capturing functional requirements in system design. Building on a user-oriented system design tool creates a natural bridge between benchmarking and validation; between the laboratory experiments for benchmarking search engines and interactive information access studies that can validate the starting points of the benchmarking studies.

When the hypotheses concerning the system usage and evaluation criteria are formulated as use cases, they can be debated and validated as well as used for setting parameters for system benchmarking. When the parameters related to users and system usage are explicitly set, it becomes easier to understand the experiments and interpret the experimental results even for stakeholders from outside the scientific community. This facilitates knowledge transfer and technology take up. It will also facilitate replication of the experiments and curation and reuse of the results by producing a descriptive layer of information concerning the experiments.

2.1 A short introduction to use cases

Use cases are a well-established system development methodology. A use case is a relatively informal or semi-formal description of a system's behaviour and usage which is intended to capture all the functional requirements of a system by describing the interactions between outside actors and the system to reach the goal of the primary actor (Jacobson 1987; Jacobson et al. 1992; Cockburn 2002; Overgaard and Palmquist 2004). In other words, in a use case a *system*, with its *primary actor* (the user), the *goal* of the primary actor, outside *actors* that the system relies on to achieve its goals, and the *sequence of actions* between the system and the actors are defined to capture and organize the functional requirements of the system. The actions of the primary actor, as formalized in the use case, are mapped onto system components and system development objects — most often using Universal Modelling Language — for the purposes of system development and evaluation.

Use cases are typically organized around a main success scenario, which describes the simplest path through the use case, the one in which everything goes right and the goal is

reached without difficulty. Also all the other scenarios, both those leading to success (possibly through recovery) and those leading to goal abandonment (failure) are described. Each scenario is an instance of the use case, a possible path through it. Usually several scenarios are needed to describe all the required system functionality (with respect to that use case). Sections typically included in use cases are show in Figure 2.1. Also additional information such as the priority and the frequency of the use case and related higher or lower level use cases may be described.

Use case name Goal Scope of the system Level Preconditions Primary actor Secondary actors Trigger Main Success Scenario Alternate Scenarios (as extensions of the main success scenario)

Figure 2.1 Sections typically included in use cases.

A well-worn out example of a simple use case is that of a cash withdrawal, where the main success scenario might be as follows:

1. Customer inserts a bank card.
2. System requests authentication.
3. Customer inserts PIN code.
4. System prompts customer to select services.
5. Customer selects withdrawal of money.
6. System prompts the customer to indicate the amount to be withdrawn.
7. Customer enters the amount.
8. System (displays a message,) ejects card and dispenses money.
9. Customer collects card and money.

Extensions are needed to handle situations such as customers entering the wrong PIN code, system running out of bills and not being able to dispense the requested amount of money and customers requesting a withdrawal not allowed by their account balance.

2.2 The implicit use case of ad hoc information retrieval

In the context of Cranfield-style information retrieval evaluation actors are typically not separated properly from the system proper that is to be evaluated. The systems are treated as black-boxes, where different components (e.g., query and document representation and

matching mechanisms, language or image processing components) are not considered as separate actors having their own use cases and deserving their own evaluations. The evaluation consists of assessing the ranked output of the system against the input request. Consequently IR evaluations produce a single figure as a result for complicated interaction effects of several components, where the gain or loss in performance becomes difficult to localize and explain. Primary actors and their goals and interaction with the system are rarely explicitly discussed in Cranfield-style studies, but are implicitly included in the experimental design as caricatures of focused, active and well-spoken users working on topical, well-defined, static and exhaustive retrieval tasks. Essentially, this kind of studies can work well to establish the usefulness of systems with respect to activities that fit this narrow use case. If the activities do not fit, evaluations will fail to establish success criteria. As information access technology has moved from this current prototypical domain of topical text retrieval, it has become less (and less) motivated to focus the research efforts on this implicit use case alone. The advent of multimedia as a large information carrier may be the most obvious example, as multimedia is different, used differently, by different users, and for different reasons than text. Thus, to capture the most important criteria for success for a variety of information access systems benchmarking should change to accommodate a variety of users with a variety of needs and goals and searching under varying conditions in varying contexts.

This is where use cases show promise of being a useful tool for evaluation of future generations of information access systems. They can be a practical tool to bridge the divide between benchmarking and validation and they can guide the design of benchmarking efforts by requiring the evaluation design to make explicit the intended usage of the evaluated system, and how it provides value for its users.

2.3 The framework

We have developed an initial framework for writing use cases for information access evaluation. The goal has been to build a resource that could support experimental design in the field of information access by making explicit the user-related functional system requirements and their connections to benchmarking mechanisms. The framework is based on the use case methodology, but the structure has been modified somewhat. All of the central components of use cases are in place, but they have been specified to a quite detailed level through identifying several features related to them that can affect the design and evaluation of information access systems. One of the strengths of the use case methodology is the low threshold of starting to use it. Use cases are rather informal and short documents. Our aim is not undermine this strength. Use cases that follow the structure of this framework do not have to be all-embracing and cover every feature defined in the framework. Rather, the framework should work as a check list over the features that potentially should be considered; or as a cook book for writing use cases that relate to some concrete information access situations, actors and goals. Not all use cases need to use all of the features. Most use cases will generate many parameters that could be set and hypotheses that could be tested. Overly complicated experiments are naturally not the goal — identifying many issues does not hinder concentrating on only some of them. Nevertheless, the different features of the use cases interact in various ways and

consequently it is difficult to give simple answers on how each feature in isolation should be evaluated. Therefore it is important to have the whole use case when designing and experiment and selecting evaluation measures.

The structure of the use case framework is presented in Figure 2.2. The framework begins with a summarizing description of the use case. After that the system features are presented, followed by features related to the primary actor. Finally, the features related to interaction between the primary actor and the system are discussed in the session features section. The features related to each of the sections are discussed below The features should guide the writing of a use case by showing what kinds of features could be considered and what kinds of values they might get. We are confident that neither the selection of the features nor the selection of the values are comprehensive. Rather it is expected that every (or most) first generation use case written using the framework should identify new features or values that could be added to the framework and the framework could be gradually extended to cover a wider selection of information access use cases than we have been able to imagine today.

- USE CASE DESCRIPTION
  - Use case name
  - Use case Summary
  - Usage Narrative
- SYSTEM FEATURES
  - System boundary
  - Secondary actors
    - Repository
    - Service provider
    - Morphology processing
    - …
  - Utilities
    - Devices
- USER FEATURES
  - Primary actor (user)
  - Task Context
  - Local context
- SESSION FEATURES
  - Goal
  - Elements of the interaction pattern
    - Search
      - Query
      - Browsing & Navigation
    - Inspect/assess
    - Manipulate
    - Export
  - Main Success Scenario
- PRECONDITIONS

Figure 2.2. Structure of the use case framework.

2.4 The features

2.4.1 Use case description

2.4.1.1 Use case name

The name of the use case should be short yet informative and should be related to the primary actor's goal.

2.4.1.2 Use case summary

The use case summary should describe what is searched for, for what reason, and how (i.e., the central features of the repository and/or the queries).

Example 1: "Find documents for supporting decision making in medical diagnosis."
Example 2: "Find an entertaining film clip with other groovy links to free sites on the same topic for me to spend the next fifteen minutes with until my bus arrives at its destination."

2.4.1.3 Usage Narrative

Usage Narrative is a situated and highly specific example of an actor using the system. It is considered to be "optional": it does not add anything to the use case really, but may be a good starting point for considering what the central issues might be. When writing the narrative, invent a fictional but specific actor, and capture, briefly, the mental state of that person, why they want what they want or what conditions drive them to act as they do.[1] Note that narrative is not a use case, but a situated scenario.

2.4.2 System features

2.4.2.1 System under discussion (SuD)

Defining the boundaries of the system is central in use case analysis. The goal is to describe the scope of the system, what it does and what it does not do. To do this, the outside actors (secondary actors) that the system relies on to achieve its goal need to be identified. Otherwise, it is easy to end up with a system that includes other systems, which in turn makes it difficult to know what should be evaluated. The system boundaries can be defined as a core retrieval engine as in the Cranfield model or more inclusively with respect to its usage and the services it provides. Correct definition of the system boundaries requires identifying the secondary actors of the system and system utilities, such as input and output devices.

2.4.2.2 Input and output devices

Identifying and describing the devices is not only important for making the system boundaries clearer, but also because the devices may have a major effect on how the system is used and for what purposes. An example of the effect of the device related features is to consider using the same system (e.g., a music distribution system) on a PC or

[1] Cockburn (2001) Writing effective use cases.

[2] This direction was taken in agreement with EPO patent examiners present at CLEF-IP 2011

using a smart phone: how will the slower input means and small screen affect the primary actor's interaction with the system and success criteria? It is possible that same tasks result in very different interaction when different devices are used. Features related to devices are described in Table 2.1.

Feature	Example values	Relation to user behaviour	Relation to evaluation
Device	Table-top, laptop, cell phone, game console, MP3 player, e-book device, TV		
Display size			
Input means	Typing, voice, pointing, clicking… keyboard, keypad, microphone, touch pad, mouse		
Output means	Text, image, video, sound, … Screen, speakers, earplugs, paper, …		

Table 2.1. Features related to devices.

2.4.2.3 Secondary actors

Secondary actors can be human or other systems that the system under discussion (SuD) needs information from to achieve its goal. Identifying secondary actors can be difficult sometimes. A good rule of thumb would be (Bittner and Spence 2003): "If you can't control it, it's an actor.". In information access evaluation the parts of the system set-up that are not the target of the evaluation and not part of the functionality that is developed should be considered secondary actors.

Repository

Repository is obviously more than "just" a secondary actor, as it is the piece that the whole system is built around, to enable access to its content. Therefore the features related to the repository will affect system design and evaluation in many ways: the typical information needs and search goals depend on the genre and content modality; the quality and credibility of the content will affect relevance criteria, as will the cost and ease accessing the information objects. All these features (and others) will potentially affect the way the system is used, i.e., typical interaction patterns. Features related to repositories are depicted in Table 2.2.

Feature	Example values	Relation to user behaviour	Relation to evaluation
Media	Text, audio, image, video, graphs, 3-D objects, maps, diagrams, structured data...		
Genre	News, entertainment, encyclopaedic and factual, personal commentary (blogs, tweets, comments), learned essays, technical text and manuals, commercial, transactional, …		
Accessibility	Unrestricted – restricted		
Provenance	Service provider/external; one source or several sources; known or unknown sources		
What is a record? What is the basic unit of content?	"Audio recording +notes + metadata", "audio only", "reference database" ...		
Document structure	Structured – unstructured		
Source dynamics: is the information source static or mutable?	Collection — stream		
Permanence of collection	None, short, long, permanent / years		
Quality of content	High – low; credibility		
Technical quality: image quality, OCR quality, …	High — low; constant — varied…		
Size	Innumerable, size (GB)		
Language	Mono/bi/multi-lingual; languages		
Indexing timeliness	Immediate, daily, weekly, monthly, …		
Coverage of the data source(s): is the information there, if you just could find it?	Complete — sampled		

Table 2.2. Features related to repositories.

Service Provider
Features related to the service provider are presented in Table 2.3.

Feature	Example values	Relation to user behaviour	Relation to evaluation
Who maintains the collection and the IR system ?	Name or purpose: BBC, public service broadcaster. Canal+, commercial.		
Business model	No cost, subscription, pay-per-view, advertisement (direct revenue/subsidized revenue/content licensing)		
Trust in service provider	Low — high		

Table 2.3. Features related to the service provider.

Other secondary actors
There are many possible other secondary actors, especially systems that the system under discussion needs to interact with to reach its goal. The secondary actors depend on the system and also on how the system boundaries are drawn – if the whole complex system is treated as a black-box and evaluated as a whole, or if the performance of a specific part-system is considered. Different secondary actor systems may have different relevant features, but generally it could be of interest to consider how they interact with the system under discussion (how often, at what point(s) in the flow of interactions) and what information the system may need from each secondary actor to function properly.

2.4.3 User features

Information needs of individuals are often described to arise from problematic situations the individuals face. This picture of information needs and information access being inherently problem-based is somewhat limited – many especially leisurely information needs are not really based on problematic situations or gaps in knowledge, as pointed out by, e.g., (Elsweiler et al. 2010). Thus “information” should include such things as entertainment – search needs that are more related to passing time or changing mood, finding some music to play at the background while focusing on other things etc. Nevertheless it is clear that individuals perceive their information needs in a highly subjective way that depends both on the context and personal characteristics of each individual (Belkin 1980; Ingwersen and Järvelin 2005). Children do not search the same way as adults do (Bilal and Kirby 2002), experts do not search the same way as novices do (White et al. 2009), and field engineers hanging from telephone poles definitely do not search the same way as people sitting in

their offices or couches at home do, as hinted by Gery Ducatel in his ECIR industry day presentation in 2011 (Ducatel 2011). Consequently, if users with their context, previous knowledge, preferences are ignored in system design and evaluation, systems cannot possibly advance beyond a particular level of accuracy on average for a specific user (Allan 2003).

2.4.3.1 Primary actor

Primary actor is the actor whose goal the system is supposed to assist. In information access context this means the person who is searching for information, who has some information need and search goal. Many different potentially important characteristics of the primary actor can be identified: demographic features such as age, gender or educational level, expertise on a domain or language skills. Information retrieval is not necessarily an individual activity performed in an isolated situation, but may be performed in a more or less direct collaboration with others: tasks can be divided between different individuals; help can be asked and received; and the preferences of the peer group might guide relevance criteria. The features related to the primary actor are presented in Table 2.4.

Feature	Example values	Relation to user behavior	Relation to evaluation
Identity: Who are the envisioned users of this system?	Genealogist, radiologist, somebody who wants to be entertained, ...		
Role	Consumer, owner, creator, editor/repository manager		
Collaboration	Single user, group		
Domain expertise	beginner, advanced beginner, competent, proficient, expert	Query formulation ability; what is relevant	
Collection expertise	beginner, advanced beginner, competent, proficient, expert		
Search experience (Routine in searching information using IR systems, NOT knowledge of the inner workings of a search system)	beginner, advanced beginner, competent, proficient, expert		
System expertise (This specific system, SuD)	beginner, advanced beginner, competent, proficient, expert		
Language skills (with respect to the information sources and task, native and foreign language)	Beginner, elementary, intermediate, advanced, proficient, native	Ability to formulate queries and assess the results? Is language support needed? Can user take advantage of language tools?	
Demographic variables	Age, gender, educational level and other socio-economic and geo-spatial variables	Example "age": What and how is seeked, for what reason? Interaction patters — all might change with age. "Children cannot read and write."	

Table 2.4. Features related to the primary actor.

2.4.3.2 Task context

The task and the domain affect the relevance criteria, search strategies and motivation of the primary actor. The primary actor is likely willing to put more effort into solving important work tasks with potentially high cost of errors than in fleeting entertainment needs. Some leisurely needs may relate to the primary actors social status and be perceived as rather important. Features related to the task context are presented in Table 2.5. In many cases the evaluation that is needed is not that with the aim of optimizing precision and recall of a standard ranked result lists, but evaluation based on an understanding of the primary actor's situation and preferences. Issues such as query formulation and browsing support, result presentation, varying relevance criteria and stringency of the relevance criteria and cost of the interaction need to be accounted for.

Feature	Example values	Relation to user behavior	Relation to evaluation
Domain	Medical, IP-chemical, Cultural heritage, entertainment, gaming, general, …		
Conventions and restrictions on the domain	Confidentiality, high data security, high cost of errors, requirement concerning the coverage (of data sets, of search results)		
Task type	Work, leisure, …		
Task	"diagnosing", "gaining social relevance", "passing time", "being entertained", "learning"		
Task complexity	Simple – complex		
Task importance	Low-high; fun-lifesaving; …		
Use case dependencies	Self-contained – part of a complex process		
Frequency of the task	One time, recurrent, routine...		

Table 2.5. Features related to the task context.

2.4.3.3 Local context

The features related to local context are described in Table 2.6. Local context is the short term, situational context of the primary actor that affects her information access behaviour. The network connection (latency and cost) affect the interaction as the primary actor tries to minimize the cost of the interaction, in money and in time. The location and position of the primary actor affects the typical information needs, relevance criteria, preferences for query

formulation and result presentation. A common example is that of searching for restaurants using a smart phone: it seems reasonable to assume that distance from the searchers location to the restaurant would affect the relevance of the restaurants. If the information need is not (yet) properly defined or focused then the primary actor might have difficulties formulating the need as a query and in assessing the relevance of documents. The stage of the search process (task) also affects the motivation, feelings and thought related to the task (Kuhlthau 1991; Liu and Belkin 2010).

Feature	**Example values**	Relation to user behavior	Relation to evaluation
Network latency	Response time		
Time constraints, task urgency	urgent –laid back		
Cost The cost of using the network, system, or service. Who's paying? Does the user care about the cost?	Low – high; irrelevant		
Motivation	Low – high		
Stage of the search process	Initiation, selection, exploration, formulation, collection, presentation (following Kuhlthau's stages)		
Goal orientation/definition of information need	Vague – "working towards a well-defined goal"		
Location, geographical	Stockholm, Egypt, South America		
Position	Home, office, train, out at the town, …		

Table 2.6. Features related to the local context.

2.4.4 Session features

Information access is inherently an interactive process and thus the interaction should not be disregarded in information access evaluation. Already Bates (1989) challenged the traditional view of information retrieval as one-shot query-result interactions with respect to static information needs. Recent studies have shown that users of information retrieval systems often search in sessions of several short queries instead of using well-defined, verbose one-shot queries (e.g. Smith and Kantor 2008). Recent studies have also shown that users can successfully compensate for the poor performance of search systems and that consequently the search engine quality as measured by MAP or nDCG correlates only weakly with task performance (Turpin and Scholer 2006; Smith and Kantor 2008).

Restricting information retrieval evaluation to one-shot queries leads to ignoring factors such as the user effort for query formulation and assessing the results. Therefore, evaluation of information retrieval systems should to be extended to include the whole sessions of user–system interaction, as suggested by Keskustalo et al. (2009).

The session features of the use case framework aim to describe the user—system interaction through a variety of possible user actions, i.e., steps in the flow of interaction. Session length is a factor that affects the interaction greatly: does the user have the possibility and/or motivation to engage in a long interaction with the system, or are results needed immediately? On the other hand, session length is a possible success criterion for a system, where typically the shortest path to satisfactory result would be preferred. The user actions included in the framework include different search strategies, such as querying, browsing and navigating; inspecting and assessing the result, saving the results and manipulating the content. These actions can be assigned different costs based on the use case: e.g., urgency, user motivation and task importance may affect the maximum cost of the session (time, effort); device may affect query formulation and result inspection cost; domain and search expertise may affect the query formulation cost, etc. and can thus help incorporating many of the use case features in evaluation

2.4.4.1 Goal

The features related to the search goal are depicted in Table 2.7. The goal taxonomy follows the taxonomy of web search described by Broder (2002) and further detailed by Rose and Levinson (2004). The taxonomy classifies queries according to their intent to three classes: navigational, informational and transactional queries. Navigational queries have the immediate goal of reaching a particular site that the user knows or assumes to exist. Thus navigational queries are known-item queries and usually have one “right” result. Informational queries have the intent of acquiring some information from one or several information objects, the goal is to learn something. Thus informational queries can be seen as type examples of the typical “information need-based” queries of standard information access studies. Transactional queries have the goal of performing some web-mediated activity, or to reach a site where some further interaction will happen. Typical categories for such queries are e.g. shopping, downloading various types of files and finding web-mediated service. The success criteria for such queries may depend on various factors that are important for the primary actor, such as price of goods, quality of content, speed of service, etc. (Broder 2002. Rose and Levinson (2004) created a web search goal hierarchy where the top level hierarchy resembles Broder’s taxonomy, except for the transactional category that has been replaced by a slightly more general category, resource queries. Rose and Levinson’s hierarchy is presented in Figure 2.3. This hierarchy gives a good starting point for considering possible user goals for information access. The goals are identified for web search and it is possible that it needs to be extended to cover other kinds of goals that occur in other search contexts (or even in web search context).

The directness of interaction depends obviously on the goal of the primary actor, but also on time constraints, on how well-defined the information need is, task stage and maybe even on the primary actor’s motivation. Search tasks oriented to task completion lead to

very different interaction patterns than open ended searching with no particular goal in mind. Very diffuse or vague information needs may also lead to more or less capricious interaction. The interaction patterns may be complex and the interaction is not always directed towards a single result. The "result" (i.e., the goal) may be very vaguely defined, such as being entertained for a short period of time, which may lead serendipitous interaction patterns.

Feature	**Example values**	**Relation to user behaviour**	**Relation to evaluation**
Initiative	Push – pull		
Type of goal	Navigational, informational, resource (can be more specific)	Will affect the interaction pattern greatly, affects the information use, …	
Type of information	Single item, data element, topic or content, factual data, monitoring data		
Directedness of interaction	Serendipitous – directed.		

Table 2.7. Goal related features.

- **Navigational:** goal is to go to a specific known website.
- **Informational:** goal is to learn something
  - Directed
    - Closed: answer to a question with a single, unambiguous answer
    - Open: answer to an open ended question, or a question with unconstrained depth
  - Undirected: goal is to learn everything/anything about a topic.
  - Advice: goal is to get advice, ideas, suggestions or instructions.
  - Locate: Goal is to find out whether/where some real world service or product can be obtained
  - List: Goal is to get the search result list itself
- **Resource:** Goal is to obtain a resource (not information) available on web pages
  - Download: goal is to download a resource
  - Entertainment: goal is to be entertained simply by viewing items
  - Interact: Goal is to interact with a resource using another program or service, e.g. weather, measure converter.
  - Obtain

Figure 2.3. The goal hierarchy by Rose & Levinson (2004), a shortened version.

2.4.4.2 Elements of the interaction patterns

The elements of the interaction pattern are described below. The elements are named after typical actions of primary actors: searching, inspecting and assessing, manipulating and exporting. The search element is further divided into "query" and "browse and navigate" elements. These elements also include a system responsibility side: each action of the primary actor has some corresponding system responsibilities: when the features of the elements (actions of the primary actor) are specified, the functional requirements of the system are specified.

Search features: querying, browsing and navigation

The features related to the different search strategies are presented in Table 2.8. There are two main categories of search types: Queries, where the primary actor explicitly and actively formulates a query and submits it to the system expecting some presentation of the result as a reply and browsing and navigation, where the primary actor follows an existing structure to access the collection. Commonly, both of these are combined and thus supporting switching between the strategies is also important.

Inspect and assess features

The effort of inspecting and assessing the results has been found to depend on for example the task stage and task type (Liu and Belkin 2010). Difficulty of inspecting and assessing the results might affect the whole user-system interaction: it might encourage the primary actor to focus on query reformulation instead of inspecting more documents, it might make the search sessions longer and it might make the primary actor less satisfied with the interaction. Thus support for inspecting and assessing the results is important and deserves its own evaluation. The features are presented in Table 2.9.

Manipulate and export

Table 2.10 presents the features for two separate categories of user actions (elements of the interaction pattern): Manipulating the collection and exporting information objects from the collection. Sometimes the users of information success systems can be involved in enriching the content of the collection: tagging or annotating or reviewing the information objects for example or discussing them. Exporting means most often saving the information objects found on the primary actor's device for future use, but might also mean printing etc. Sometimes also information related to the search might be saved.

Feature	Example values	Relation to user behavior	Relation to evaluation
Supported search strategies	Simple query, advanced query, command-based, browsing and navigation support, …		
Support for changing between the types of search	None (start over), supported		
Query modality	Text, image, video, audio … mixed		
Query formulation	Specification, example		
Query target	Content, metadata/description		
Query support	Spelling correction, synonyms, support for advanced query language (advanced query fields)		
Navigation support	Classifications, thesauri, search result, sitemap, FAQ		
Sorted by entity	People, countries, subject, media, period, date, language, collection…		

Table 2.8. Search features: support for different search strategies.

Feature	Example values	Relation to user behavior	Relation to evaluation
Result presentation	Single-item, answer, summary, list, browsing interface, notification		
Result organisation	By score, date, diversity, …		
Granularity of result presentation	Title, snippets, thumbnails, keywords, item (text document/image/video…)		
Support	Highlighting keywords, relevance scores, showing relations within a document, …		

Table 2.9. Support for inspecting the results and making relevance assessment.

Feature	Example values	Relation to user behavior	Relation to evaluation
Level of engagement			
Type of contribution	Tagging, annotation, commenting, discussing, creating temporary lists of documents (baskets), …		
Type of access			
Save	Full context of search, queries, how many documents found, documents, sets of documents, baskets, …		

Table 2.10. Features related to manipulating content and exporting information objects.

2.4.4.3 Main success scenario

The typical interaction patterns vary greatly between different use cases, depending on the tasks and goals of the primary actors (and potentially on several other factors). A navigational web search might have the following main success scenario:

- User types in a query.
- System returns a ranked list of results.
- User clicks on the first result

For many other use cases, any reasonable main success scenario might be difficult to recognize. This is in part due to the berry-picking behaviour of the searchers described by Bates (1989). As discussed above, the interaction patterns may be complex and the interaction may be capricious or serendipitous. Many use cases could probably be compressed into "one-shot to success" scenarios, but this will essentially make the main success scenarios quite uninformative and uninteresting, just repeating the traditional exclusion of interactivity from evaluation. Therefore, the goal ought to be understanding and communicating the typical patterns of interaction through describing realistic main success scenarios together with their most central extensions.

An example of a main success scenario
1. Primary actor enters the query page and enters a query: an example image. (Primary search strategy is querying, system needs to have query functionality for content-based image retrieval by uploading an image)
2. System presents a result: 20 most relevant image thumbnails per each result page arranged by date and with links to the original image with some metadata. (Result presentation requirements based on some hypothesis on use preferences?)

3. Primary actor inspects and assesses the first result page with image thumbnails. (Assessing images is fast)
4. Primary actor clicks the link to image 3. (Browsing and navigation functionality is required)
5. System presents the original image with copyright, date and a caption, as well as some saving functionality.
6. Primary actor inspects and assesses: Takes a closer look at the image, checks the metadata. Makes positive assessment: this picture is just what the primary actor needs.
7. Primary actor saves the image.
8. Use case ends (success – primary actor just needed this one image).

Extensions:
1.b. Primary actor does not have an example query image. Text search is also supported — primary actor types in a free text query.
4.b Primary actor does not see anything relevant. (Jump to 1 (new example image, textual query — how is query modification supported?) or end)
6.b The image is not relevant. (Jump to 3, 1 or end)
8.b Primary actor needs more images. (Jump to 1 or 3)

2.4.5 Preconditions

What is already the state of the world, what needs to be true for the use case to happen. For example Primary actor has a means of accessing the system under discussion.

2.5 Relating aspects of test collections to use case features

The use case framework describes a variety of features that can be taken into account in evaluation experiments. While Cranfield-style test collection based evaluation is just one type of evaluation experiment, it is a very successful and dominant one, and many of the evaluation tasks described in this deliverable are rooted in this tradition. This is why in this section we list some central aspects of test collection based evaluation experiments, and relate them to features from our use case framework. Not all experiments require describing all of the use case features – only the ones relevant to the experiment need to be considered. Typically, it is useful to consider how the use case affects the most central parts of the test collections, i.e., document collection, search topics and relevance assessments. Some considerations are summarized in Table 2.11 below. So far in the deliverable the main focus and discussion has been on arriving at a relatively stable and complete, easy to use framework to describe use cases. Table 2.11 can help in further discussion on how we should specify test-collection based evaluation tasks in such a way that we demonstrate to what extent we are measuring user satisfaction.

Part of the evaluation setting	Corresponding use case features	Values
User Model	Primary actor, goal, local context, session, …	The model for the user interaction with the system: time/effort for different actions; user preferences. User, task, context etc. related features that affect the previous, e.g., the location and position of the user.
Test collection	Source	Modality, size, quality, dynamics, business model – cost affects relevance criteria (of real users), etc. The realism of the collection vs. availability and control of the experiment.
Topics, work task descriptions	Task, goal	Work tasks, leisurely tasks, entertainment. The complexity or simplicity of the tasks. Frequency of the task. Informational, navigational and resource goals
Relevance assessments	Task, goal, …	Binary - graded. Expert – layperson. Algorithmic, topical, situational, emotional, socio-cognitive, … criteria: source, recency, structure, quality, difficulty, style, coolness, famousness, language, genre, modality,
Interface	Search strategies	None, simple, complex, search box, browsing interface, combined, …
Queries	Queries	Explicit, implicit, short, long, well-defined, ambiguous, automatic, manual, one-shot, sessions, …

Table 2.11 Some central parts of a test collection and their connections to the use cases.

3 Use cases and associated evaluation tasks

In this section we work out use cases from the three main domains under study in the PROMISE project: the “search for innovation” domain, the medical domain and the cultural heritage domain. In addition, we discuss two use cases from another domain: entity search or more specific: people search. Each domain has characteristics that play a role in use cases from that domain. The level of discussion in this chapter however is not the domain level. Each section discusses a single use case, and for each use case the associated domain is indicated. At the end of each use case section there are one or more sections on evaluation of the use case. There can be one or more evaluation tasks designed for evaluation of different subsets of the desired functionality expressed in the use case. In some cases evaluation tasks do not quite fit with the use case. For example, the “Variability”

task in Section 3.3 is one of the tasks that will be run as part of the CHiC lab at CLEF 2012. Rather than targeting the "Search for lecture material" use case, it targets a more general cultural heritage use case. At the time of writing of this deliverable, the design of evaluation tasks in the cultural heritage is still in full swing. In the next deliverable we aim to describe use cases for each evaluation task.

3.1 Visual clinical decision support for medical diagnosis

The task we study in this use case is to find medical cases/images similar to the one under observation for supporting a clinician's decision making during medical diagnosis using medical images and text describing the case under observation as queries in biomedical literature. To get an idea of a typical situation, we describe a hypothetical scenario now. After that, we discuss system features, user features, session features and evaluation in some detail.

Usage narrative

Alonso, a medical graduate, is currently a second year intern in the radiology department of a large university hospital. The clinician supervising Alonso has asked him to perform a medical diagnosis on a patient and has provided him with the patient's latest MRI scans and medical record. Unable to reach a decision as he is not 100% sure about the diagnosis and potential co-morbidities, Alonso decides to search the literature for similar cases by using as queries the MRI scans and also text that describes the medical case under observation. A successful end would be for Alonso to find articles in the literature that help him decide on a medical diagnosis.

As part of his training, Alonso has become quite familiar with medical cases and images, but he does not have yet substantial experience in searching the PubMed Central collection for locating similar cases in the literature. He has used search systems before (e.g., the Web search engines), but he has no knowledge of the internal techniques of IR systems (i.e., he is IR illiterate). Although his mother tongue is not English, his language skills allow him to formulate English queries.

System features

The System under Discussion (SuD) is a biomedical literature retrieval system.

The platform being used can be a desktop or a laptop or a tablet computer, or a cell phone without display size restrictions. The input can be provided through typing or clicking, while a keyboard, mouse or touchscreen would be ways for interacting with the system.

The repository, a biomedical literature collection such as, e.g. PubMed Central, typically contains millions of scientific articles published in biomedical journals and other venues such as conferences and workshops. These articles are mostly written in English and a large number of them contain images and graphs. They are high-quality and trusted sources since they are peer-reviewed with a known provenance. Such collections are updated in regular intervals (e.g., weekly) with timely additions of recent scientific articles and are

expected to be maintained for the foreseeable future. Their coverage of the literature published in the field is generally very comprehensive.

Such collections and retrieval systems are typically maintained by organizations that provide access to biomedical libraries and tools, such as National Center for Biotechnology information (NCBI) of the National Library of Medicine (NLM) in the USA. These are highly trusted service providers that follow a no cost business model.

User features

The primary actor is a clinical practitioner searching the biomedical literature to find information relevant to a medical case under observation on the basis of the patient's medical imaging exams and medical record; this primary actor has the role of a "consumer" of the information access system.

The primary actor is typically a single user with a higher level of education, but with varying levels of domain and collection expertise (ranging from medical students and interns to Professors of Medicine) and also of system expertise (ranging from novices to clinicians with significant experience in using such medical information retrieval systems). However, the primary actors have no knowledge of the internal techniques of IR systems, i.e., they are IR illiterate. Furthermore, the language skills of the primary actor with respect to the information sources, i.e., the biomedical literature which is mostly written in English, are typically at the very least adequate and very often excellent. Finally, the demographic variables cover a wide spectrum in terms of age (ranging from young medical graduates to older experienced clinicians) and of socio-economic and geospatial variables (ranging from a clinicians working in a small hospital in a rural area to those employed by a large university hospital in a metropolitan area).

The task context for this use case is a medical domain. Since the information sources used are scientific articles published in the literature, there are no confidentiality issues. The database is potentially accessible to all clinicians.

During their daily work routine, the clinicians need the information access system to decide on a medical diagnosis for a specific patient given the patient's medical exams and in particular medical images and the overall medical record. This is a complex task since there is a large amount of information to handle and there is also a need to work with multimodal information (structured data, text and images). This task is highly important as it can be life-saving for the patients under observation.

The clinicians are highly motivated to use the system because the online access is free and the system supports them in their decision making. The response time should be fast, as clinicians should find relevant information quickly to prevent frustration and time-loss. The typical location is a hospital during daily clinical routine. However, since online access is

provided, clinicians may further use the system after work to continue their research on the particular case.

Session features

The goal is clinical decision support for the medical diagnosis of a specific patient under observation on the basis of evidence from their medical imaging exams and medical record. This is an informational task where the aim is to get advice, ideas or suggestions from scientific articles describing medical cases similar to the one under observation and containing images similar to the ones from the current case.

We now look at elements of the interaction patterns relating to searching, the queries, browsing and navigation, inspecting and assessing results, and exporting or saving searches and / or results. Then there follows one concrete example of a successful interaction with the system.

The main type of search is querying through either a simple or an advanced query interface. Support for browsing and navigation should also be provided, together with support for changing between the different types of search.

Queries are formulated both through specification and also through providing examples and include multiple modalities (structured data, text or images). Advanced query support features improve the effectiveness of the performed searches. See Figure 3.1 for an interface used in this use case domain. Here a physician is performing diagnosis and the system highlights regions that were automatically classified into specific patters. The next step of the person will then be to search for similar cases. This is an example of a multimodal query with a complex structure.

Navigation support can be performed through filtering the search results (or even the whole collection) based on various features, e.g., the modality acquisition of medical images, the patient's age and sex, and also metadata, e.g., the author names, journal titles, or MeSH terms (Medical Subject Headings) of the articles in the biomedical literature.

Search results are presented as a list sorted by relevance. It is desirable for each result to be presented with its title, a snippet with some text relevant to the query (possibly with the query terms highlighted), and thumbnails of the images it contains. Additional information, such as the MeSH terms under which it is classified or the number and types of images it contains, can also be displayed.

Saving past queries, possible with the whole list of results, or individual search results would be desirable.

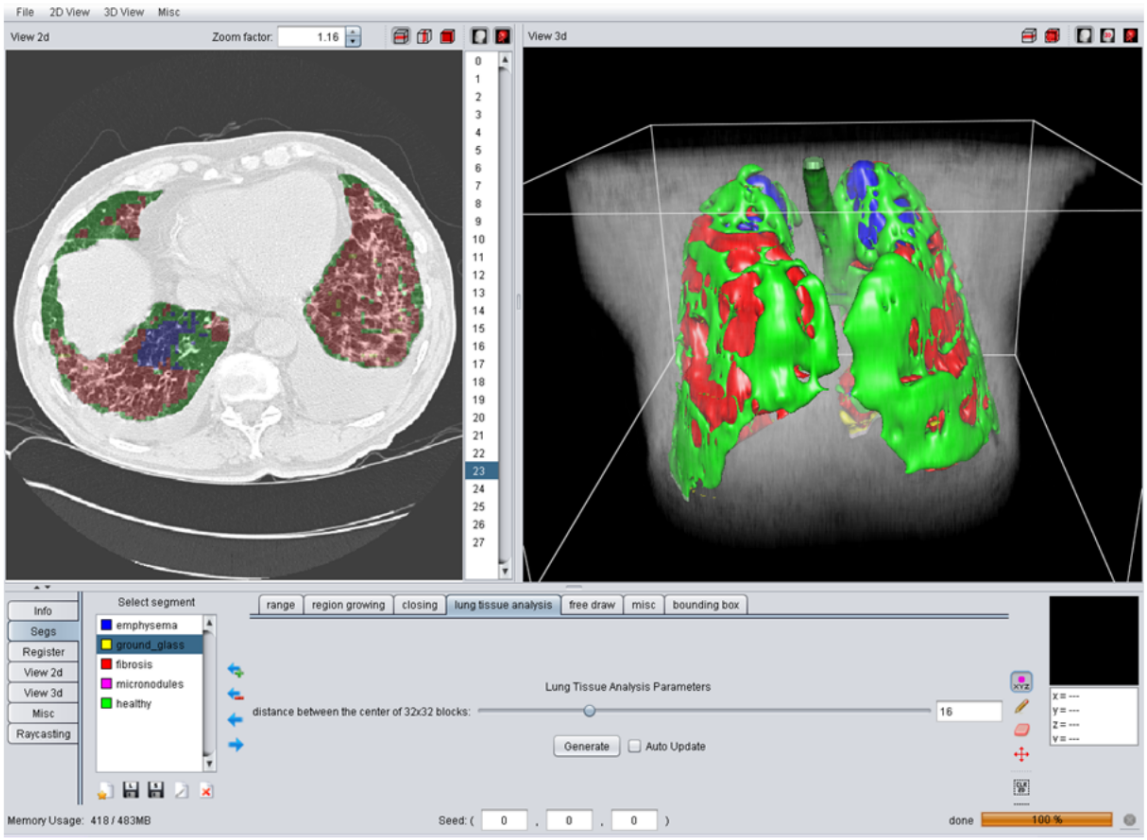

Figure 3.1: a user interface used in the medical use case domain. A physician is performing diagnosis. The system highlights regions that were automatically classified into specific patterns. The next step of the physician will be to search for similar cases.

One example of a successful flow of interaction

1. The clinician chooses to use the biomedical search engine to find similar cases to the one he is diagnosing.
2. The clinician formulates the query (using text, example images or regions, structured data).
3. The system retrieves the results according to the defined criteria.
4. The clinician peruses the first result page and clicks on few of the results to read the articles in more detail.
5. Every time a result is clicked, the system presents the full article from the biomedical literature together with its metadata and the images it contains.
6. END: success, the clinician finds the images and articles that help him make a decision on the medical diagnosis of the case under observation

Two medical evaluation tasks: image-based and case-based retrieval

An evaluation task based on the visual clinical decision support for medical diagnosis use case should evaluate several aspects: effectiveness remains the most important, together with efficiency, whereas evaluation of the usability of the user interface to maximize the clinicians' satisfaction with the full system should also be considered. Currently, the medical task at ImageCLEF focuses on the evaluation of the effectiveness of medical case and medical image retrieval, with MAP being the main evaluation metric, while the evaluation of user-oriented aspects (such as usability), e.g., through an interactive medical retrieval task, is among our goals.

The medical image-based task has been running at ImageCLEF since 2005. The focus is on the retrieval of similar images for a precise information need. In 2009, the retrieval of similar cases was introduced. The goal of the case-base retrieval subtask is to retrieve cases including images that might best suit the provided case description.

The topics are information needs in three languages and images. They are developed based on the query logs of a web-based image retrieval system that provides medical professionals access to radiology resources and are subsequently validated by a clinical practitioner (typically a radiologist).

For the image-based retrieval task, textual queries (i.e., mega cisterna magna) with some sample images for each query were given to the participants. In contrast, for the case-based retrieval, a case description, with patient demographics, limited symptoms and test results including imaging studies, was provided (i.e., A 63 year old female remarked an un-painful mass on the lateral side of her right thigh. Five months later she visited her physician because of the persistence of the mass. Clinically, the mass is hard and seems to be adherent to deep plane).

The test collection has grown from 8,000 images in 2004 to over 230,000 images in 2011. To ensure the high quality of the test collection, the documents are obtained from highly trusted sources of biomedical literature, such as PubMed.

Due to the unfeasibility of manually reviewing all images for all topics, "pooling" was used to reduce the number of candidate images for each topic to ~ 1000. The limitation of this method is that the pools used for relevance judgments reflect the runs submitted by the participants. Therefore, images that may have been retrieved by other techniques or were not the top hits would not be evaluated. Relevance assessments are performed by expert users, clinicians or medical students.

To evaluate the submissions of the participants, the standard TREC evaluation software (http://trec.nist.gov/trec_eval/) is used (trec_eval). The results are calculated using the following standard measures given by trec_eval: MAP, P10,P20, Rprec , bprec, bpref and num_rel_ret.

3.2 Prior art search

The task in this use case is to find documents describing products or methods similar to those described in the document at hand for assessing the innovative step of a patent application or invention disclosure using the text of the application or queries derived therefrom in patent and non-patent literature. Before we describe system, user and session features and evalution in detail, we give a narrative of a hypothetical but typical scenario.

Narrative

A professional searcher, Ginés, receives a document describing a potential invention. If Ginés works for a patent office, this document is a patent application document. Otherwise, this is generally a less formal document, a so-called innovation disclosure. In either of the two cases, Ginés' task is to identify other documents describing similar products or methods, and to assess, based on these documents, the innovative step of the invention. For this, he connects to one or more repositories of information and issues text queries. In some cases, he would prefer to issue non-textual queries, based on the images in the patent application document, but the systems used either do not provide such a feature, or the feature does not provide useful results.

After each query, he investigates the list of results, marks some for future reference, and possible searches again, with a different query.

In the end, he creates a list of documents and marks them as being highly relevant or simply relevant and creates a search report to send back to the applicant (or invention discloser) and to keep on record.

Ginés is fluent in at least two languages, has a comprehensive technical vocabulary for a specific field in at least these two languages and is confortable issuing queries in English.

System features

The System under Discussion (SuD) is a desired, hypothetical but realistic, system.

Ginés accesses it with desktop or a laptop for his professional activity. He may receive a request (an application or innovation disclosure) also as a printout and may also have to submit his report as a printout.

The repository is a collection of patent and/or scientific articles, consisting of text, images, graphs, 3-D objects, diagrams, specific data formats (e.g. chemical entities) and specific metadata (IPC classification, inventors, applicants, dates, identifiers). In Figure 3.2 we see an example chemical image. The text of the collection is generally of a scientific nature, with inserts of a legal nature. The repository has restricted access, even though some of its parts are public information.

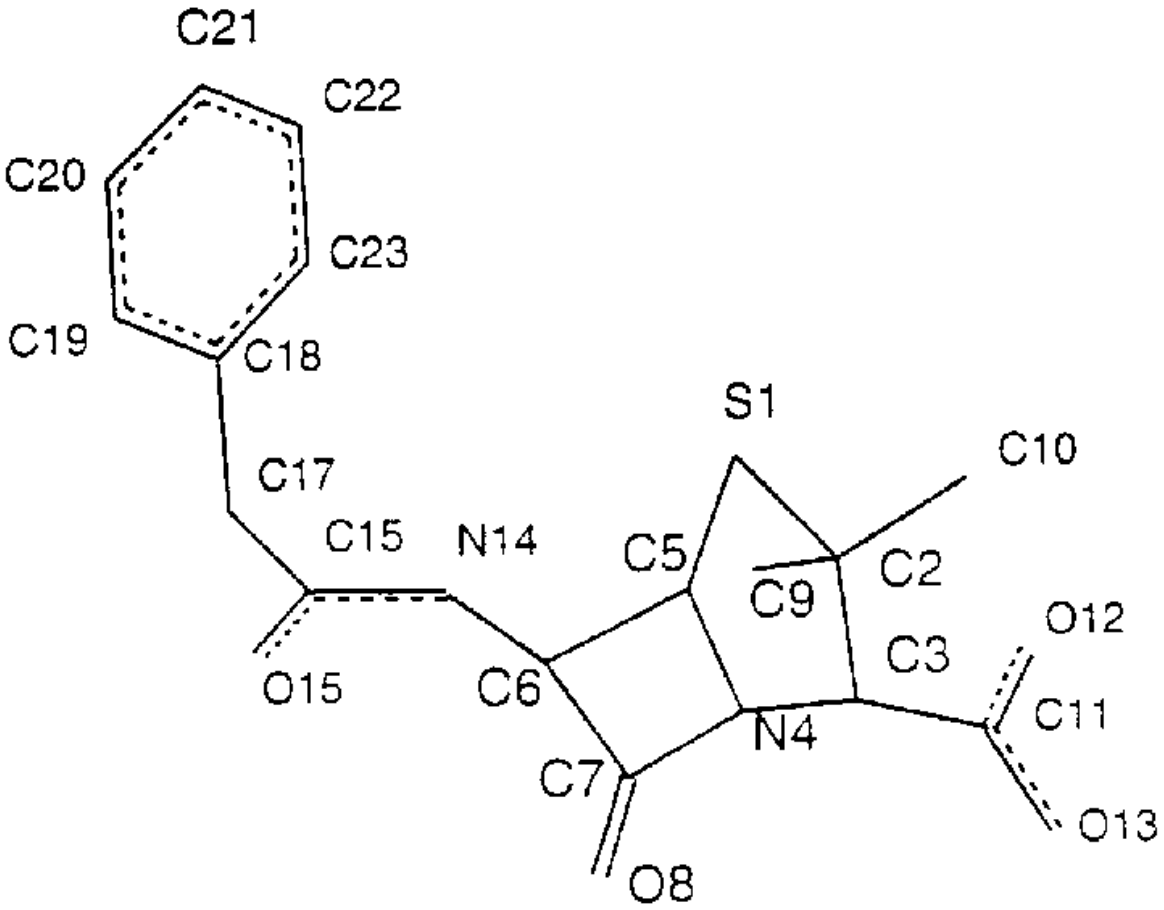

Figure 3.2: Example of chemical image from a repository in the search for innovation domain.

The basic unit of content is a structured document and a document, once published, does not change. Instead, amendments and changes are done through issuing other documents. While the expected quality of each document is high, the technical quality may vary (sometimes documents are scanned copies of hard-copies). The size of the repository varies between hundreds of GB to several TB and generally covers documents in several languages. The timeliness of the index varies, function of the source of the document. The aim is to index as soon as a document becomes available. The coverage is complete in the sense that it is generally known what is in the repository.

The service provider is generally a commercial or governmental entity, highly trusted, paid either by subscription or subsidized by a government or supra-governmental entity.

User features

The primary actor is a professional searcher, Ginés. He works for a patent office, a consulting firm, or a large corporation with an R&D department. His task is to find relevant documents for a patent application or innovation disclosure. He is the consumer of the repository, but also a contributor. Although he is an expert in his field, Ginés may on occasions collaborate with colleagues to create a final list of relevant documents. In addition to expertise in his field, he is also an expert in search methods and is extremely familiar with the system he uses.

Ginés speaks at least two languages fluently, is able to issue queries in all of them and to interpret results in all of them, even though he would welcome the assistance of an automated translation system. Ginés is an adult with higher education in a specific field. No other demographic specifics can be identified.

The task domain is intellectual property, of various technical fields. Ginés has an important, complex work task, generally confidential and with potential high cost for errors. The task, with different search objects, is recurrent, at a frequency of 1-2 per day. It does not directly depend on other tasks.

The task is under moderate to high time constraints. The cost of performing the task can be quite high, function of the repository used, and Ginés is moderately-to-highly motivated in finding all the relevant documents. The environment of the searcher is an office.

Session features

The goal of the search, according to the Rose & Levinson (2004) hierarchy, is Information/Directed/List. We now discuss elements of interactions related to searching, queries issued, browsing and navigation, inspecting and assessing results, and exporting or saving search sessions and results. Then there follows one example of a successful flow of interaction.

The system must support simple and advanced queries, filtering, browsing and navigation support. It must allow the user to switch between search and exploration modes.

The queries are given in text form, but this is more due to the limits of the system than to the utility of other media. The system may have a range of query support functionality (spelling correction, synonyms, etc.) but Ginés must remain in control of what query is actually used in the search process.

The system allows the user to filter on classification, dates, issueing authority (in the case of patent documents). It also supports sorting on a variatey of fields.

Search results are presented as a list, preferably with snippets that allow Ginés to see why a document was retrieved. Additionally, the snippets have highlighted keywords from the

query. They are sorted per computed relevance score, but additional information is clearly visible in the search result list (date, other versions, publisher).

Ginés is not allowed to manipulate any document in the repository. He creates his own document, based on the list of results.

Ginés would like to be able to save any used queries and respective found documents, as well as any document baskets he created

One example of a successful flow of interaction

1. Ginés receives a patent application document to evaluate
2. Ginés enters the search system
3. Ginés enters a text query, potentially with 40odelle operators and specific field filters.
4. System present a result list, sorted by relevance, with snippets, metadata information, and links to full documents.
5. Ginés inspects and assesses all the documents
6. Ginés clicks on one element of the list for further inspection.
7. System presents the full document, with any metadata, attached images and text.
8. Ginés inspects and assesses. Finds the document potentially relevant and saves it to a bucket.
9. Ginés clicks on the « Back » button to return to the list of results.
10. System presents the list, with the already viewed documents visibly identifiable.
11. Jump to Step 6.
12. Based on a new understanding, Ginés inputs a new query
13. Jump to Step 3.
14. Ginés considers the result list he identified as sufficient
15. Ginés saves the list and creates a search report
16. Use case ends.

Evaluation tasks for the prior art search use case in CLEF-IP

The CLEF-IP evaluation campaign has been, and thanks to PROMISE, continues to be, the instantion of this use-case in a systematic evaluation exercise. The 2009-2011 Prior Art Candidates task has focused on Step 3 of the above example of a successful flow of interaction. It has used as topics application documents, as collection a large set of European and International patent documents, and as measures a range of metrics to account for the different understandings of *sufficient* in Step 14.

In addition to the text query, Step 3 also involves specific field filters. Very often, such filters refer to the International Patent Classification (IPC) class of the patent, a widely used and simple classification scheme for patents. The 2010-2011 CLEF-IP Classification tasks have

looked at the ability of a system to support the search by automatically identifying the class of the patent application at hand, or of the innovation disclosure document. As for the Prior Art Candidates task, the topics here were patent application documents. The measures considered were Precision and Recall at 1 and 5, and a combination thereof (F1 measure). The measures are indicative for the two types of filters used: strict (the top class should be used) and fuzzy (up to 5 classes accepted). Arguably, the number 5 is arbitrary in this case, but nevertheless indicative of the fuzzy filter scenario.

As mentioned in paragraph 3 of the Session Features section above, the user may find it more useful to use the images present in an application document, rather than the text. Step 3 does not currently describe this feature because there are no systems available to provide this feature at a sufficiently high level of quality. It is one of the secondary purposes of our evaluation campaign to encourage research groups to work on issues which would make the example scenario above more efficient and successful for the user.

This is why, in 2011, the CLEF-IP had two image related tasks, in addition to the two text-only tasks just described. The first image related task, image-based retrieval, corresponds to the Prior Art Candidates task but requires participants to also use the images in the application document to find relevant prior art. The second task evaluates a smaller, useful functionality that could be part of a prior art search service: classification of patent images in classes such as abstract drawing, graph, flow chart, gene sequence, and so on.

After the first three years, where the step was taken as a whole (i.e. given a document, present a list of documents), the 2012 instance moves deeper, to the features of the result list. In particular, we look now at a particular type of search in Step 3 and at Step 8 (document assessment). Namely, the evaluation focuses on how well can the search system identify the most relevant paragraph(s) to show the primary actor, in order for him to take a decision[2]. We crystallize this in two tasks:

1. **Passage retrieval starting from claims (patentability or novelty search)**: The topics in this task will be based on the claims in patent application documents. Given a claim, the participants will be asked to retrieve relevant documents in the collection and mark out the relevant passages in these documents.
2. **Matching claim to description in a single document (Pilot)**: Given one claim in a patent application document, the participants will be asked to indicate those paragraphs in the description section of the same application document that best explain the contents of the given claim.

Task 1 is a particular type of search at Step 3. Using a claim as a starting point is, according to the users we interviewed, a common process in the case of novelty search, resulting from the nature of the patent document. The claim is the part of the document, which explicitly lists the exact method or entity for which protection is being sought. As such, while the description section of the patent application may contain many items not novel in any way,

[2] This direction was taken in agreement with EPO patent examiners present at CLEF-IP 2011

the claim is the essence of the innovation. The tasks is that much more difficult as the language of the claim is an unnatural version of human language. The results will be judged, in the first phase, based on the full documents retrieved by the systems.

A second evaluation on Task 1, together with task 2, target Step 8 of the example successful flow of interaction above. This is the step where the user must analyze the document and make a decision with regards to its pertinence to the application document or innovation disclosure at hand. As many patent documents are tens of pages long (with extremes going into thousands of pages), some systems highlight what they consider to be the most important paragraphs. Such paragraphs can be in other documents (Task 1) or in the same document (Task 2).

A bit more background on Task 2: The problem modeled here, as described by professional users, is that the claims are sometimes written in such unnatural language, that it is hard even for an expert to understand what exactly it is referring to. In such cases, practice indicates that in order to understand the claim, the examiner has to look at the description of the invention, and search those paragraphs that provide context to the claim at hand.

At the time of writing this deliverable, the measures to be used for these tasks are still under investigation.

As mentioned in paragraph 3 of the Session Features section above, images play an important role, which currently is under-used due to the inability of the systems to semantically process images. The difficulty of processing patent images is notorious and has been documented before [Hanbury et al, 2011]. In 2012 we organize two tasks to better understand the problems and to encourage research in the area:

3. **Flowchart Recognition Task**: The topics in this third task are patent images representing flow-charts. Participants in this task will be asked to extract the information in these images and return it in a predefined textual format.
4. **Chemical Structure Recognition Task**. The topics in this fourth task will be patent pages in TIFF format. Participants will be asked to identify the location of the chemical structures depicted on these pages and, for each of them, return the corresponding structure in a MOL file (a chemical structure file format).

Task 3 above will allow the user to input a more precise query at Step 3, using the text inside flowcharts and even the relationships between flowchart nodes. The data used is representative for the problem: a collection of flowcharts from patent documents. The evaluation will be done manually to both achieve a high degree of confidence in the results and a better understanding of the shortcomings of automated systems. The metrics, while still under investigation at the time of the writing of this report, will involve both the identified graph structure and the text therein.

Task 4 is a highly demanded feature. As a corpus we will use 500 pages selected by a large pharmaceutical company, Astra Zeneca. Results will be evaluated manually. The measure used is the number of chemical compounds correctly identified. The difficulty here is in judging what ‘correctly identified’ means, due to the specific nature of the corpus. Experts in the field are part of the organizing team.

3.3 Search for lecture material

The task in this cultural heritage domain use case is to find Second World War pictures for a history class presentation using textual queries in public sources of cultural heritage collections such as digital libraries. We assume an educational user type and a search for objects relying on an available persona developed by EuropeanaConnect [Guldbæk Rasmussen et al. 2010]. Again, we first give a fictional narrative to give an idea of typical use before we discuss system, user and session features and evaluation.

Narrative

Ricote, history professor, searches in a digital library in order to prepare a presentation for his lecture. For this reason he needs a selection of various images in high-definition of soldiers. He decides to use Europeana as a portal to access material from libraries and archives that are under no legal restrictions. Starting with textual queries he investigates the result lists and saves successful queries in his user profile for future research. From the result list he browses through the “explore further” function to find similar images to the one that were already displayed according to his search queries. Finding a matching image he uses an out link to the content provider in order to view and safe the original object. In the end he creates a collection of images. In Figure 3.3 we see an example of a matching image.

Ricote has active language skills in at least 2 languages as well as excellent domain knowledge and can easily browse through results in several languages.

System features

The System under Discussion (SuD) is an information system accessing cultural heritage material (images) such as Europeana. Ricote uses a desktop or laptop at his office.

The repository is a large-scale reference database dealing with metadata as basic units and providing linkages to the original content in external institutions like libraries, museums or archives. The documents are highly structured and available in different European languages as well as in various media types, like text, audio, image or video files. The collection level depends on the participating institutions, but is well organized and under current enrichment. The cultural heritage is public by definition, therefore, the accessibility is not restricted and Europeana in general is a non-profit organization maintained and promoted mainly by the European Commission. The technical quality and the trust are high and the provenance has an excellent reputation.

The service is provided and maintained by a non-profit organization maintained and promoted mainly by the European Commission.

Figure 3.1: screenshot of Therese assessing a search result.

User features

The primary actor Ricote is a professional searcher, working at the university. He takes is a single end-user with an higher educational level, excellent language skills and a background as middle aged, female, upper class US-American. He is an expert in the domain of cultural heritage, using databases and/or portals like Europeana occasionally, his information retrieval literacy is high and his user mood is based on professional routine.

The task represents a use case from the Cultural Heritage domain. The task is carried out in a scientific setting, related to the daily work of Ricote as a history professor at the university. It requires high quality data for educational purposes.

The task is carried out under moderate time constraints with a well defined goal. Ricote works from his office at the university and either accesses free available databases or uses the university license for restricted repositories.

Session features

The goal of this task is an overview / list of images related to the topic for further use. According to Rose & Levinson (2004) the goal can be classified as: informational/directed/list. We now discuss elements of typical interaction patterns relating to searching, queries, browsing and navigation, inspecting and assessing results. Unique in SuD is that it facilitates some manipulation of the collection, and saving of a search.

The system supports at least a simple search function as well as filtering, browsing and navigation functionalities. The combination of search and browsing actions should be easy and intuitive.

Currently only text base queries are support but other input forms are desired, e.g. uploading an image to find similar material, or even humming to find the Turkish March. The system offers different support functionalities such as spelling corrections, disambiguation and auto-completion.

The system allows filtering of the search results via facets such as media type, provider, language, country, date and copyright. It also supports similarity search based on a search result returned for a previous search query.

Results are shown as thumbnails and can either be displayed as a list, sorted by media type or through a timeline. The full result display provides extended meta data information about the object as well as the link to the original object itself. Meta data information can be translated into the preferred language.

Users can tag and annotate found objects.

The system provides a function to embed the retrieved object as well as social media sharing functionalities. Via the user profile search queries as well as objects can be saved. Ricote would like to prepare a presentation directly supported by the system.

One example of a successful flow of interaction

The main success scenario could be described as the following basic flow of interaction: Ricote selects Europeana as portal to different collections of cultural heritage objects and enters the query “world war II soldiers”. After receiving a result list he is browsing through the first result pages and refines the results according to format, date and subject. Subsequently he clicks on a few thumbnails to find appropriate images and leaves the portal through an outlink to the content provider. According to the use case framework the interactions after the outlink and further usage (e.g. saving, writing, transforming, adopting, annotating, merging) of resources are not considered but could have some relevance for information retrieval behaviour.

1. information need triggered by work task
2. selects Europeana as portal to different collections of cultural heritage objects
3. enters the query: “world war II soldiers”
4. gets back a result list
5. browsing through first result pages
6. refines results according to format, date and subject
7. clicks on a few thumbnails to find appropriate objects
8. eave the portal through an out-link to content provider
9. comes back to result page
10. back to 5
11. clicks on a full result view
12. uses browsing functionalities to find similar objects
13. clicks on another result
14. safes some objects and search terms in his user profile
15. creates a list of images
16. end of use case

Preconditions

Ricote must be aware that the system he uses only provides metadata and is limited to European Cultural Heritage objects.

Evaluation in the cultural heritage domain

Due to the diversity of material, the visual discovery options and flexible access requirements expected in cultural heritage information systems, functionalities such as time line and map-based browsing, virtual exhibitions and discovery through social features are equally important as searching in this context (Minerva Working Group 5, 2008).

Two important considerations for evaluations in the cultural heritage domain that deserve separate mentioning are the availability and willingness of test subjects for user-centric studies and the difficulties to develop standard test collections and create relevance assessments for system-centric approaches. Due to the variety of users, finding a representative user sample is labor-intensive, time-consuming and – in the case of multicultural and multilingual cultural heritage systems – crosses country and language borders. Traditional IR test collections have been built over many years and often with institutional support (e.g. NIST for TREC). Several cultural heritage systems have served as test cases, for example in the TEL (The European Library) track at at CLEF (Agirre et al., 2009; Ferro & Peters, 2010) or the INEX book track (Kazai & Doucet, 2007; Kazai et al., 2010).Building a cultural heritage collection often not only deals with many different content types, but also with many different rights owners. Clearing (rights) and cleaning (format)

cultural heritage data for standard, comparative evaluation initiatives has been proven a complicated undertaking.

In line with the CLEF2012 conference, the CHiC 2012 pilot evaluation lab will support a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. Data test collections and queries will come from the cultural heritage domain (in 2012 data from Europeana) and tasks will contain a mix of conventional system-oriented evaluation scenarios (e.g. ad-hoc retrieval and semantic enrichment) for comparison with other domains and a uniquely customized scenario for the CH domain. We now describe three planned cultural heritage evaluation tasks for CLEF 2012, showing that considerable progress has been made in the specification of evaluation tasks in the cultural heritage domain since deliverable 2.1 (Karlgren et al, 2011). The specifications of the tasks may still see some changes before they are run at the CHiC 2012 lab at CLEF 2012.

The ad-hoc task at CHiC

Task definition: This task is a standard ad-hoc retrieval task, which measures an information retrieval effectiveness with respect to user input in the form of queries. No further user-system interaction is assumed although automatic blind feedback or query expansion mechanisms are allowed to improve the system ranking. The ad-hoc setting is the standard setting for an information retrieval system - without prior knowledge about the user need or context, the system is required to produce a relevance-ranked list of documents based entirely on the query and the features of the collection documents. For CHIC, it will also serve to develop a baseline for system performance. We will test monolingual, bilingual and multilingual retrieval in 3 major European languages: English, French and German.

Collections: The collection used is Europeana (www.europeana.eu), a large digital library, museum and archive, which provides access to over 20 million cultural heritage objects. The documents in the Europeana collection are metadata records consisting of brief descriptions of the object (title, keywords, description, date, provider) and occur in multiple languages. For experimental purposes, the Europeana collection will be divided into 3 subcollections according to metadata languages, so that some control over the language of documents for the relevance assessments can be asserted.

English collection: all Europeana documents with English metadata records.
French collection: all Europeana documents with French metadata records.
German collection: all Europeana documents with German metadata records.
Europeana collection: the complete Europeana collection with all metadata records.
More detailed specifications on the collections will follow once the collection is released.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. Navigational queries are rarely seen in Europeana, however queries for people, places and works (named entities) occur very often. The 50 short topics in title-format only (e.g. “Eiffel tower”) reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.) in a cultural heritage digital library researched previously.

Expected results: Participants are expected to submit relevance-ranked result lists for all 50 topics in TREC-style format.

Relevance assessments: Relevance assessments will be done manually by first collaboratively generating an assumed information need for the query and describing it (which will be used for later editions) and assessing the pooled documents for their relevance according to the query + information need. This assumes the perspective of an average user (we assume the majority of users typing that particular query would have that particular information need).

Evaluation metrics: The evaluation metrics for the ad-hoc task will be the standard information retrieval measures of precision and recall, particularly the standard measure mean average precision (MAP) and precision@k.

The 'Variability' task at CHIC

Task definition: This task requires systems to present a list of 12 objects (represents the first Europeana results page), which are relevant to the query and should present a particular good overview over the different object types and categories targeted towards a casual user, who might like the "best" documents possibly sorted into "must sees" and "other possibilities." This task is about returning diverse objects and resembles the diversity tasks of the Interactive TREC track or the CLEF Image photo tracks. For CHIC, this task resembles a typical user of a cultural heritage information system, who would like to get an overview over what the system has with respect to a certain concept or what the best alternatives are. It is also a pilot task for this type of data collection. We will test monolingual, bilingual and multilingual retrieval in 3 major European languages: English, French and German.

Collections: The collection used is Europeana (www.europeana.eu), a large digital library, museum and archive, which provides access to over 20 million cultural heritage objects. The documents in the Europeana collection are metadata records consisting of brief descriptions of the object (title, keywords, description, date, provider) and occur in multiple languages. For experimental purposes, the Europeana collection will be divided into 3 subcollections according to metadata languages, so that some control over the language of documents for the relevance assessments can be asserted.
English collection: all Europeana documents with English metadata records.
French collection: all Europeana documents with French metadata records.
German collection: all Europeana documents with German metadata records.
Europeana collection: the complette Europeana collection with all metadata records.
More detailed specifications on the collections will follow once the collection is released.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. The 25 topics reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.) but will be enhanced with query categories that show different ambiguous aspects of a topic (e.g. topic = "Chardonne", categories: person, place).

Expected results: Participants are expected to submit 12 relevant results for all 25 topics in TREC-style format. Documents should be as diverse as possible with respect to:

- media type of object (text, image, audio, video)
- content provider
- query category
- other features to be described / suggested by participants

Relevance assessments: Relevance assessments will be done manually by first collaboratively generating an assumed information need for the query and describing it (which will be used for later editions) and assessing the pooled documents for their relevance according to the query + information need + variability / diversity. If possible, we will compare 2 types of assessments: cultural heritage experts vs. "naive" users of cultural heritage information systems in order to be able to compare their assessments of relevance and variability.

Evaluation metrics: The evaluation metrics for the variabililty task will be the standard information retrieval measure of precision, particularly the standard measure mean average precision (MAP) and precision@k as well as diversity measures used in the Interactive TREC track like cluster-recall and intent-aware precision, which might be adapted to the diversity requirements set forth in this task.

The 'Semantic Enrichment' task at CHiC

Task definition: The task requires systems to present a ranked list of at most 15 related concepts for a query to semantically enrich the query and / or guess the user's information need or original query intent. Related concepts can be extracted from Europeana data (internal information) or from other resources in the LOD cloud or other external resources (e.g. Wikipedia). Semantic enrichment is an important task in information systems with short and therefore ambiguous queries like Europeana, which will support the information retrieval process either interactively (the user is asked for clarification, e.g. "Did you mean?") or automatically (the query is automatically expanded with semantically related concepts to increase the likely search success). For CHIC, this task resembles a typical user interaction, where the system should react to an ambiguous query with a clarification request (or a result output as required in the variability task). We will offer the task and topics in 3 major European languages: English, French and German.

Collections: The collection used is Europeana (www.europeana.eu), a large digital library, museum and archive, which provides access to over 20 million cultural heritage objects. The documents in the Europeana collection are metadata records consisting of brief descriptions of the object (title, keywords, description, date, provider) and occur in multiple languages. For experimental purposes, the Europeana collection will be divided into 3 subcollections according to metadata languages, so that some control over the language of documents for the relevance assessments can be asserted.
English collection: all Europeana documents with English metadata records.
French collection: all Europeana documents with French metadata records.
German collection: all Europeana documents with German metadata records.
Europeana collection: the complete Europeana collection with all metadata records.

For semantic enrichment, the Europeana Linked Open Data collections can be used: Europeana released metadata on 2.5 million objects as linked open data in a pilot project. The data is represented in the Europeana Data Model (RDF) and encompasses collections from ca. 300 content providers. Other external resources are allowed but need to be specified in the description from participants.
More detailed specifications on the collections will follow once the collection is released.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. The 25 topics reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.).

Expected results: Participants are expected to submit 15 ranked different terms or phrases for all 25 topics which express semantic enrichments for the query in the respective language and could be used for query expansion.

Relevance assessments: Relevance will be assessed in 2 phases:
(1) First all submitted enrichments will be assessed manually for use in an interactive query expansion environment (e.g. “does this suggestion make sense with respect to the original query?).
(2) The submitted terms and phrases will be used in a query expansion experiment with a standard IR system, i.e. the enrichments will be individually added to the query and submitted to the system. The results will be assessed according to ad-hoc retrieval standards.

Evaluation metrics: The evaluation metrics for the semantic enrichment task will be the standard information retrieval measure of precision (+precision@1 and @3) for the first phase of assessing just the submitted enrichments and the standard ad-hoc information retrieval measures for the second phase of assessing the submitted enrichments as query expansion variations.

3.4 Expert profiling in a knowledge intensive organization

Expert finding is a well known and well studied problem (Bailey et al, 2007; Balog et al, 2009, Craswell et al, 2006). Balog (2008) also studies the closely related task of expert profiling. Expert finding answers the question ‘Which experts know about X’?, while expert profiling answers the question ‘What topics does expert X know about’?. Using the use case framework, in particular it’s example of a successful interaction, it will become clear how the two tasks are related. The expert finding and profiling tasks are not in the three main PROMISE domains. Instead, the general domain for these tasks is entity search, or people search in particular.

We study the task in the context of a knowledge intensive organization (e.g a university). Our approach is to rank expertise areas from a thesaurus of expertise areas for a given expert using his name and any known associations with documents in a collection of documents such as scientific publications, supervised master theses, course description

and homepages. While the findings about what kind of algorithms are suitable to tackle this task should apply to other knowledge intensive organisations, we obtained ground truth from a particular organization, the university of Tilburg. Before we turn to evaluating systems, we describe the use case using the use case framework.

Usage narrative with a fictional user and system

Sancho wants to profile an expert, Alfonso, from a knowledge intensive organization . It could be that Sancho is an employee at this organization and wants to consider if Alfonso is suitable for a meeting with a visiting employee from another organization. Sancho may also be somebody from another organization who is interested in Alfonso, e.g. to consider if he would be a suitable candidate for a current job opening. Sancho's task is to report back or take a decision based on Alfonso's areas of expertise. Sancho knows that Alfonso's organization has a representation of the collective knowledge of their employees in the form of a thesaurus of expertise areas. There is also a search engine in which the name of an employee can be entered, upon which a ranked list of expertise areas is returned. Sancho wishes to use this search engine to get an initial idea of Alfonso's expertise. After that, he may use other search engines, e.g. Google Scholar to gain more information. In the end, after possibly consulting Alfonso, or people who know Alfonso, Sancho reaches a decision. NB: a closely related use case is expert finding: given an expertise area, Sancho wants to rank the experts based on how proficient or knowledgeable they are in this area. We will see how the two use cases can be alternated in a single sitting, with one user goal in the basic successful flow of interaction below.

System features

A knowledge intensive organization, e.g. a university makes available the system under discussion (SuD). The system accepts as a query an expert name. It ranks for this expert a list of known expertise areas, which are organized in a thesaurus, or knowledge base. To rank the areas depending on the expert, the system has available an intranet-like repository collected and maintained by the knowledge intensive organization. It may be publicly accessible or it may not. Explicit associations between experts in the organization and the documents in the collection may or may not be available. The collection and the knowledge base may be multilingual. The collection will be continuously growing.

User features

The primary actor, Sancho in our narrative example, is usually an adult, searching in a professional setting. He usually has at least a concept of the scope of knowledge in the organization. He knows Alfonso already, or he knows of the existence of Alfonso, or he found Alfonso in a prior expert finding query. In any case, Sancho now wants to profile Alfonso.

The task the primary actor faces is to get an idea of what Alfonso's areas of expertise are and to what extent Alfonso is an expert in them. If Sancho is part of support staff, the task may be recurrent. If Sancho is interested in hiring Alfonso, the task is more incidental. The

task is complex. Ranking expertise areas from the knowledge based of the organization is only a first step for Sancho to get a realistic idea of Alfonso's areas of expertise. There are other factors that play a role in the decisions Sancho will make, such as social factors, availability of Alfonso, etc. See (Hofmann et al, 2010) for a study of contextual factors in expertise seeking. There are other experts that Sancho might be interested in instead of Alfonso.

Time constraints, Sancho's motivation, cost of errors (missing areas, unrelevant areas) will vary with the exact nature of Sancho's task, and the decision he will take based on the results. For example, is he assessing the suitablitity of Alfonso to have a productive meeting with visiting researchers, or is he considering to invite Alfonso to apply for an open full professor position?

In the first case, Sancho may base his decision on the output of the expert profiling system. In the second case, Sancho would definitely consult many other sources of information as well.

Session features

The goal of Sancho in a single sitting is to find the areas of expertise of Alfonso. It could also be that profiling Alfonso is a sub-goal in an expert finding session. In this case Sancho could have started with a query consisting of an expertise area from the knowledge base, and found Alfonso among the top experts. By clicking on Alfonso, he would trigger an expert profiling query to the system. Conversely, if, in Alfonso's profile, Sancho clicks on an expertise area, he would be issuing expert finding query.

An expert profiling query can be given in text form, e.g. "FirstName LastName". We assume that ambiguity of names will not be a major problem in a moderate size knowledge intensive organization. If uncertainty exists as to which expert was intended, this could be solved by presenting a list of matching experts: clicking on one of these experts would then result in an unambiguous query. In the Rose & Levinson (2004) hierarchy an expert profiling query would be Informational / List.

The result list of expertise areas should be clickable, each area should link to a detailed description of the area. The detailed description of the area should contain a list of other related areas, broader and narrower or otherwise. Ideally, a click on an expertise area would also result in an expert finding search: A list of experts ordered by the degree to which they are an expert in this area would then also be shown on the detailed area page. Then, if the expert that Sancho is currently profiling (e.g Alfonso) is not in the top ranks of this list, Alfonso should still show on the result page with his rank on this area in brackets behind his name.

A successful basic flow of interaction

Sancho receives visiting researchers from oversees and wants to make sure that they have a good time here. He also wants to make sure that the visiting researchers meet the best people from Sancho's university on the areas of interest of the visiting researchers. Sancho his first thought is Alfonso, even though Sancho has only a vague idea of the general area Alfonso works in. To get a more fine grained idea, Sancho fires up a web browser, navigates to the home page of his university and from their to the expert finding / profiling system.

1. Sancho enters the query 'Alfonso'.
2. He receives a result list of expertise areas, ordered by degree of expertise in it of Alfonso.
3. Sancho thinks some of the top areas could be of interest to his visiting researchers.
4. Sancho clicks on one of these top areas (Thereby performing an expert finding query!)
5. Sancho reads the detailed description of the area and sure enough it is relevant.
6. Sancho sees that Alfonso ranks as one of the top experts in this area. (in the expert finding result list)
7. Optional: Sancho navigates to some related areas.
8. Optional: Sancho navigates to some other top experts.
9. Sancho navigates back to Alfonso.
10. Jump to 4.
11. Sancho knows enough, Alfonso is the perfect man to meet his visiting researchers.

An evaluation task for expert profiling

For the evaluation of expert profiling systems, we have a test collection. We now relate the choices made here to properties of the use case described above. The test collection was released in 2007 (Balog et al, 2007), and we are working on releasing an updated version of the test collection that will have a new set of relevance assessments obtained by asking experts to judge profiles we generated for them. This work is supported by PROMISE.

Our test collection is designed for evaluating a specific component of the above use case, namely the quality of the ranked list of expertise areas that Sancho obtains in Step 2 of the successful flow of interaction above. We describe properties of the test collection, the relevance assessments, and some simplifying assumptions that are made to arrive at a benchmarking evaluation task.

First of all, the task assumes that a ranking of expertise areas is sufficient to assist Sancho in the task of expertise profiling. This is a simplification. This can be seen by considering the case that Alfonso is no expert in anything. A system could then return assign a zero score to each expertise areas and return expertise areas in an arbitrary order. Ideally, systems would also take a binary decision for each area whether or not it should be included in the profile.

Our repository and knowledge base of expertise area was obtained from Tilburg University. A previous version of this repository and knowledge base has been used in previous work, e.g. by Balog (2008). While the real repository is dynamic--new publications are incoming on almost a daily basis--our test collection repository consists of a snapshot of a university corpus as described. We store only textual content, even though images are also part of the repository. It is a publicly accessible corpus maintained by a university and consisting of employee homepages, scientific publications, supervised student theses and course descriptions. The size of the corpus is approximately thirty thousand documents. The corpus is bilingual (English and Dutch). To accommodate the fact that some users are more fluent in Dutch while others are more fluent in English, the test collection allows for the evaluation of English as well as of Dutch queries. The organization has a thesaurus of expertise areas (the knowledge base of areas). It consists of roughly 2.500 expertise areas. It is also bilingual (English and Dutch). Expertise areas have relationships between them: 'Broader then', 'Narrower then', 'Preferred term', 'Related'.

As in Cranfield style experiments, our relevance assessments are relations between a query (an expert), a document (an expertise area), and a relevance grade (level of expertise). So far, the level of expertise has only been binary, but in our update of the test collection there will be graded relevance assessments as well. The test collection contains relevance assessments performed by the experts themselves. This is a special situation compared to Cranfield style experiments, where assessors judge documents for queries they are not necessarily an expert on. In our setting, the query coincides with an expert, and this expert herself provided the assessments. Thinking about our users again, we can interpret this as follows: We assume our users trust the experts in their self-proclaimed expertise. Another way to look at it is to say: We assume that our experts--who were aiming to create a coherent and concise profile when they selected areas for their profiles--, are representative for real end users and their information needs. Indeed, users are looking for a complete and concise profile of experts. However, the functionality that has been tested so far is ranking expertise areas, and a long ranked list is not the same as a concise profile: this is a slight mismatch.

For evaluation of the ranked lists we use standard measures for information retrieval, such as MAP to assess the quality of the entire ranked list and MRR as a measure of early precision. Standard evaluation measures only take into account query-document relevance relations. This means that the relevance of an item to a query is independent of the relevance of items that were retrieved at a higher rank. However, experts have indicated that they do not appreciate overlap in their profiles. To capture this aspect, De Rijke et al, 2010, propose a metric that penalizes redundancy and rewards near misses. We are considering to study this metric in future work. Another interesting direction for future work is to use the test collection for a slightly different task: estimate the level of expertise for all tasks, possibly in comparison with colleagues in the same organization. This would require systems to do more than just ranking areas, and such output would perhaps be of more use to end users.

3.5 People searching for people that have been in the news.

This use case is not developed here as an evaluation task. To develop the use case framework, it is beneficial to describe many use cases with it, including more experimental, less common ones, even if it is unclear how to frame them as evaluation tasks yet. We include this use case for the validation of our use case framework, then, and not for the validation of an evaluation task.

The 'People searching for people who have been in the news' use case is inspired by a log analysis of a people search engine (Weerkamp et al, 2011). In that log analysis, it was found that some person names are searched for by many different people during a short period of time, while before and after this period the name is rarely searched for. A sample of a few thousand query instances from the logs was annotated by hand, and almost five percent of them were classified as 'event based queries', queries for which the information need was likely related to somebody who played a role in a recent event, such as a car crash, a murder, or a talent show finale. Of these five percent, 33 percent were related to deaths, 23 to criminals, 10 percent to celebrities, 10 percent to other high profile people, 9 percent to television, and 6 percent were sex related.

In summary, the primary actor in this use case is searching for information about somebody who has played a role in a recent event, usually covered in the news. The query will be a person name, possibly with a keyword such as a location. The source collection is the internet. The system under discussion (SuD) is a people search engine. It is a meta search engine that queries social media platform search engines and general purpose web search engines. There now follows a hypothetical example usage narrative.

Usage narrative

Aldonza is reading news headlines on a national news server and reads that a young man named S. van den Berg from Amsterdam was killed in car incident. She feels a sudden pang of sorrow, because she remembers going to high school with a Simon van den Berg who later on moved to Amsterdam to study there, at which time they lost contact. Anxious to find out if it is indeed her high school friend that passed away she visits a people search engine, and enters his full first and last name. Van den Berg is a very common name in The Netherlands and the SuD offers many social media profiles from national and international social media platforms. Aldonza frantically clicks on all top ranked profiles that could belong to her friend, to find out if there could be any information available. The SuD also lists documents that were returned by major web search engines. Aldonza scans the result snippets for information about a recent car accident and hopes to find one source that will state the full first and last name of the deceased.

System features

SuD accept queries of the form "Firstname Lastname". In an advanced search interface that is a bit hard to find, it offers an additional search field where a keyword may be entered, which is done in about four percent of the queries (Weerkamp et al, 2011). The system

queries a number of social media platforms and a number of general purpose web search engines. It also lists result lists for various kind of media, such as documents and images. In addition, it lists 'tags' associated with this person name.

SuD is not in charge of the repository, it does not do indexing. It is a metasearch engine. It does do some caching. In scope, the repository is all data available on the Internet. However, since it is known that a query will be a person name, the search engine can mine documents or other items for specific information. For example, it receives social media profiles from social media platforms, and it will try to extract a user profile picture, a year of birth and so on from this profile for the result snippet. While we focus on news related queries, the SuD does not know the intent of an incoming query. However, since people searching for people who have been in the news is a common use case, the SuD could mine person name occurrences in news items.

User features

The primary actor could be anybody, as in general web search. The task context may also vary from establishing if somebody you know may have passed away to getting the latest gossip. The local context can also vary. Most searches are issued during working hours (Weerkamp et al, 2010). Some of these searches may be part of some professional tasks, others may be part of a work break: we do not know. In most cases, however, we are looking for documents about a particular person, who was in an event at a particular time in a particular place. So there are important temporal and spatial factors about the query and the information need.

Zooming in on the most common use case, death related event based queries again: Aldonza from the use case narrative is certainly highly motivated. In terms of Kuhltau's stages (REF), her search is probably in the collection stage: she is trying to get all documents relating to the accident and all the documents relating to her high school friend Simon with the specific goal in mind of finding a document about both the accident and her friend or else to rule out this possibility. We can say she works towards a well defined goal. Because an event based search is related to a certain event, for most primary actors the goal will be reasonably well defined.

Session features

In a single sitting, the goal is to 'pull' information about a particular person in relation to an event. It is an informational goal. The type of information sought will be factual in cases like those of Aldonza. In other cases, the type of content could be very different, for example photographs from a recent photoshoot of a celebrity. In Rose & Levinson's hierarchy, Aldonza her search is Informational, Directed.

Turning to the SuD, its search result page (SERP) has several tabs (panes). The first result pane shows social media platforms. A second pane (hidden behind the first) shows web search engine results. For each platform or engine a separate result list is shown. The

primary actor may expand each list by clicking on an 'expand' button. Each list has snippets, which may be expanded by another click to show more information. Finally, in the expanded snippet users may click to leave the search result pages and go to an external sites. Only these final clickouts are in the transaction logs. The effort required to do a clickout is one explanation of why in these logs we observe less clicks then in web search (Weerkamp et al, 2011).

Basic flow of interaction

1. Aldonza enters the query 'Simon van den Berg'
2. SuD returns twenty profiles for all social media platforms, with links to 'more results from this platform' at the bottom of ch platform list.
3. Aldonza expands the top platform list.
4. Aldonza inspects the mini snippets, but finds no information to rule out any of these platforms.
5. Aldonza starts expanding snippets.
6. Aldonza sees a user profile picture for the third hit. It has been a while since she has seen Simon, but this might be him.
7. Aldonza clicks on the outlink in the expanded snippet to visit the profile.
8. The profile is private, unfortunately, and there is no additional information.
9. Aldonza now follows outlinks to other Simon's profiles, even if there is no user picture.
10. Aldonza does not find her high school friend.
11. Aldonza now sees that there are also web search engine results in a second pane.
12. Aldonza starts scanning Google search results.
13. The news story is mentioned in one of them, but there is no full name.
14. Frustrated, Aldonza starts wondering if there is an Advanced Search option.
15. After some digging, she finds the option to add a keyword to the search.
16. Aldonza issues the query 'Simon van den Berg, Amsterdam'.
17. Jump to 3.
18. Aldonza finds a profile from a Simon van den Berg from Amsterdam. He is her high school friend. He still is friends with some other people from their school.
19. Aldonza decides to try and get in touch with some of them via the social media platform, still afraid that it may be her high school friend who was killed in the car accident.

Preconditions

There has been an event in the real world that is the direct cause for this search. A person name in this event is the query.

Considerations for evaluation

In Aldonza's use case scenario documents about the car accident are relevant. Documents about her high school friend are also relevant. These persons need not be the same person. In other cases, only documents about the person in the event may be relevant, for example in the use case scenario where photos from a particular photoshoot are required.

We can evaluate other aspects than can be computed from a fixed set of relevant documents per topic. For example, did SuD detect that the query related to a (recent) event (Berendsen et al, 2011)? Did it signal occurrences of the name in recent (or even old) news items and bring these to the user's attention? Did it simply rank these news items higher, or did it extract relevant passages, adding a different result pane to the search result page? Also, SuD may try to cluster search results referring to the same person together. SuD could point out social media profiles that are connected to news stories, for example. If SuD makes such attempts, simple ranked list evaluation is no longer adequate.

3.6 Historical newspaper search

In the above use case descriptions, we merged some sections from the use case framework to improve readability. In this last use case, we leave the heading structure from the use case framework intact. This way, for each aspect in the framework we have at least seen one worked out example. The historical newspaper search use case is a use case in the cultural heritage domain, one of the three main domains the PROMISE project focuses on. At the end of this section, we describe an evaluation experiment that showcases the usefulness of the use case framework. The experiment is innovative: simulation is applied to generate sessions with query modifications. The experiment is also validated in the sense that many use case feature values are directly accounted for in the evaluation setup. In addition, query terms for query modification simulation are obtained through a user study involving users from the targeted population of end users for systems under scrutiny. In short, this evaluation setup extends the Cranfield methodology with its simulation methodology, and with its extensive efforts to evaluate according to targeted end user preferences.

3.6.1 Use case description

3.6.1.1 Name

Find historical newspaper articles for academic course work.

3.6.1.2 Summary

This use case describes explorative search in a historical newspaper archive with the goal of finding source material for a bachelor's thesis. The historical word forms and variants that are frequent in the collection are unfamiliar to the primary actor and the newspapers are OCR scanned in mediocre quality. The primary actor will keep searching until the right sample of interesting and/or topically relevant documents is found, or until the time or the primary actor's patients runs out.

3.6.1.3 Usage narrative

Third year history undergrad Ana Félix is attending a seminar for writing a bachelor's thesis. She has selected the topic "high school education of women in Finland during the 19th century" and has just started working. She doesn't know yet what or how much information can be found on the topic, so she is exploring the different sources of information. One of the sources she is searching is the Finnish National library's digitized collection of 19th century Finnish newspapers. She has never used the collection before, but it has been recommended by her supervisor and is accessible free of charge over the web from any computer. So, Ana Félix consults the help page and checks the classification categories for suitable entry points. She decides then to use the free text query interface. She is interested in anything and everything related to her topic: both news articles reporting concrete happenings and opinionated articles debating the topic from different viewpoints are interesting. Ana Félix tries several short queries, reformulating them as she learns from the results: changing to better keywords and adjusting the topic. Finally she learns that a suitable topic might be "Opinions for and against founding the first gymnasium for girls in Helsinki" and she figures out a few good query formulations for finding relevant documents. She saves a whole bunch of documents for closer inspection and finishes her search session when she feels that she has enough material to work on, for now.

3.6.2 System features

3.6.2.1 System under discussion

The system under discussion is the information access system of the historical newspaper archive. This use case describes a system where the document and query representation and matching algorithms are included in the system, but with morphological processing, cognate matching and translation tools seen as secondary actors: they do affect the system performance, but are not the target of the evaluation. Query formulation and modification are covered by the system, as well as inspecting and assessing the results (by necessity as the goal of the experiment is to study the most effective query modification strategies).

3.6.2.2 Input and output devices

It is either the primary actor's private computer (laptop/tabletop) or a computer in a university computer classroom (tabletop). Thus the display size also varies, but is reasonable (no cell phone use etc). Input devices are mouse and keyboard; the means are typing and clicking.

3.6.2.3 Secondary actors

3.6.2.3.1 Repository

The digital historical newspaper archive contains high resolution scanned images of some 1,7million newspaper pages from newspapers published in Finland during 1771-1910 together with the OCR scanned text of the articles. Access to the collection is unrestricted: anyone can access the collection from any computer, free of charge. The documents in the collection originate from a variety of historical newspapers, most of which do not exist anymore. The content is not selected by the service provider: all newspapers published in Finland during the time period covered by the collection are included. The use of the

collection as a historical source requires the primary actor to understand newspaper genre in general and the differences between 19th century and contemporary newspapers (reliability: the amount of opinionated text, hearsay etc).

A record in the collection contains a PDF image of a complete newspaper, several related text files each containing the OCR scanned unstructured content of one article from the newspaper (automatically separated into files) and some metadata. The collection is intended to be permanent. It is very stable, new material is added only rarely as the digitization progresses to cover more recent newspapers. The page images in the collection are high quality. The text files on the other hand are corrupted with many OCR errors. The collection is bilingual Finnish and Swedish as the newspapers are indexed in their original language. The language in the newspapers is old and differences in vocabulary, orthography and somewhat even in morphology and syntax compared to modern Finnish and Swedish occur. Especially the Finnish spelling (and vocabulary) was only standardized during the 19th century and thus quite notable dialectal differences occur in the different newspapers. As a consequent of this and the OCR errors occurring while scanning the collection, the rate of graphical variants of a word can sometimes be very high.

3.6.2.3.2 Service provider

The service provider is the National library of Finland. It is a public service cultural institution with high reputation and credibility. There is no business model — the service is publicly financed for the benefit of “the general public”.

3.6.2.3.3 Statistical Stemmer

A statistical stemmer is used to stem both index words and query words.

3.6.2.3.4 Fuzzy matching

Approximate string matching is used when matching query words against index words to make it possible to also find historical and OCR variants of the query words.

3.6.2.3.5 Dictionary-based Query Translation

As the collection is bilingual, some query translation support might be offered.

3.6.3 User features

3.6.3.1 Primary actor

The primary actor is a history undergrad: a young person with higher education and low income. The primary actor works usually alone. The primary actor is competent in the historical domain, but a beginner in research-like work tasks; is competent in general IR system use (as academic studies require frequent use of different IR systems), but a beginner or advanced beginner in using the SuD and the collection. The primary actor is typically a native Finn with one of Finland’s two official languages as the native language. The skills in the second native language vary, but are typically advanced beginner to intermediate level: the primary actor can read the second native language (with some difficulty), but might have trouble formulating queries using it. The collection is bilingual Finnish and Swedish and thus the primary actor could be said to have native and intermediate language skills. Some language support might be needed.

3.6.3.2 Task context

The use case is placed in academic domain. The primary actor is an advanced undergrad student carrying out a work task: Writing a bachelor's thesis with the aim of learning about the good research praxis and conventions and gaining a degree. Typical requirements and conventions concerning academic work prevail, related to source quality and source criticism, referencing conventions, requirements of originality of the work. The organizational culture at universities is slightly hierarchic and very individualistic — bringing the task into a successful end is almost solely the primary actor's responsibility and interest. The task is highly important to the primary actor. It is a complex learning and writing process. This leads to complex search tasks — the type of search needs and behaviour changes depending on the stage of the information search process. Frequency: rare — recurring.

3.6.3.3 Local context

The network latency varies, as the primary actor's location may vary between home and university computer classroom. Typically, the task is not urgent and does not have to be completed on one sitting. That said, the primary actor still experiences some kind of a time pressure that may lead to limiting the task and not following too difficult leads: if a search is too difficult and not giving results, the primary actor might give it up or change direction. The primary actor is on the exploration stage of the search process and thus the goal orientation is also rather vague: the information need is not yet very well specified. The primary actor has been instructed to use this collection and is thus motivated to use it. If facing a lot of trouble when searching the collection, it is possible for the primary actor to also completely give up on using it.

3.6.4 Session features

3.6.4.1 Goal

The primary actor is active and has an informational and undirected goal of finding everything and anything about a topic. The type of information searched for is single items in their original context (for history researcher the context is also important). The vague, explorative information need reflects on the directness of the interaction that is rather explorative, or general purpose: it is not know what kinds of documents will be relevant so different kinds of items that are somewhat likely to include relevant information are consulted.

3.6.4.2 Elements of the interaction

3.6.4.2.1 Search: Querying and Browsing and navigating

Both simple query interface and some browsing support are provided. Support for switching between querying and browsing is not provided. The PDF images are retrievable trough textual keyword queries targeting the OCR scanned text and using publication dates. Spelling and OCR error correction is included. All articles are accessible through two navigation support resources: a "topical article index" originating from the end of the 19th century and newspaper name index.

3.6.4.2.2 Inspect and assess features

Results are presented as a browsing interface, where the results are organized by their score or by date. Each result consists of publication date, the name and the volume of the newspaper and a text snippet with highlighted query words as well as a link to the page image containing the complete newspaper.

3.6.4.2.3 Export features

It is possible to save the PDF-file containing the image of the newspaper.

3.6.4.3 An overly simplified example of a main success scenario

1. Primary actor navigates to the search page and types in a query into the query field.
2. System (does a lot of processing) and returns a browsing interface of ranked results including a snippet of the text where the query word(s) occur and a link to the document (the image file).
3. Primary actor inspects the snippets and the newspaper names. Primary actor clicks on a link to see the PDF document containing the complete article in its original context.
4. System presents the requested PDF and provides tools for reading and saving the text: zooming in and out, flipping page, save-button, as well as a way to navigate back to result list.
5. Primary actor inspects the document and saves a copy (as it seems relevant).
6. Primary actor is satisfied with the document found and decides to end the search session. Use case ends (success).

3.6.4.4 Extensions

3.b Primary actor does not see anything relevant. (Jump to 1, user frustration increases)
3.c Primary actor does not understand the result shown (quality of the OCR scanned text in the snippets is very low etc), but proceeds as usual. (user frustration increases)
5.b Document not relevant. Primary actor navigates back to the result list or the search page (jump to 1 or 3, user frustration increases).
6.b. The primary actor wants to find more documents. (Jump to 1 or 3.)

3.6.5 Preconditions

Primary actor has been accepted to a course for writing a bachelor's thesis and has a deadline assigned by somebody else for when the work needs to be finished. The primary actor is encouraged by an outside authority ("the supervisor") to use this source of information. The primary actor will need to present the work to her fellow students and will receive a grade for the work which increases the motivation to do a good work.

3.6.6 An evaluation experiment with query modification strategies

This use case could give rise to many different evaluation set-ups. In the following, an experiment with the goal of learning about effective query modification strategies for exploratory search in historical collections containing OCR scanned text is described. To be able to study the effect of different query modification strategies, an experimental setting

based on simulated search sessions (e.g. Keskustalo et al. 2009) including several query modifications is used, instead of just evaluating the system based on one-shot queries. The effect of each query modification on the search result can thus be measured, as well as the cumulated final result of the search session. Evaluation is based on the quality of the results and the cost of the session.

The test collection used is a typical test collection for information retrieval laboratory evaluations. It contains 180.486 (844 MB) OCR scanned Finnish newspaper articles from the 19th century, along with 56 topics (with title, description and narrative) and relevance assessments for those topics. The collection also contains "alternative titles" for each topic that record additional possible search terms for the topics. The relevance assessments are on a four point scale: non relevant, marginally relevant, relevant and highly relevant and were created by research assistants, who were not specialists in the domain of the collection. The relevance assessors were instructed in a usual manner to think that they were looking for information for writing an essay on the topic. Consequently, one might say that the collection is built from the perspective of a non-expert user working on a rather demanding writing task.

While the task of writing an essay works very well for the use case at hand, the expertise level of the assessors differs from the expertise level of primary actor (a history undergrad) considered in the study. It is known from previous studies that expertise affects relevance assessments: laypersons tend to be more lenient than experts in their assessments. Such a mismatch in relevance assessments might obviously affect evaluation results. The compromise is accepted simply because it would be unpractical to make separate relevance assessments for every user group one might want to study. On the other hand, it is also known from previous studies that experts and laypersons tend to agree on the order of documents when making relevance assessments, even if they draw the boundaries between the different relevance grades differently. In other words: if presented with two documents, an expert and a layperson would usually agree on which of the documents is more relevant, even if they would assign them in different relevance categories. Thus the stricter relevance criteria of (semi-) expert users can be accounted for by downgrading the relevance of the documents in the relevance corpus. In this study, marginally relevant documents are excluded from the relevance corpus (i.e. treated as irrelevant), which leaves us with a three point relevance scale: irrelevant, marginally relevant and relevant.

To further adjust the test setting towards the envisioned user group, the query terms used in the simulated queries will be collected from a user study involving a group of history undergrads (instead of using the topic words directly). The test subjects will be first asked to suggest query terms based on a topic description only. Then the test subjects are shown the top 10 results for their query including source name, publication date and a snippet for each document (but no access to the full text of the documents or further documents). They are then asked which documents they believe might be relevant, if they think that the query worked well and finally to suggest new query terms to improve the previous query based on the results. This step can be repeated several times, if the test subject is not satisfied with

the results. The goal of this study is to collect query terms for the simulation, but also a few open ended questions concerning the test subjects' perceptions of the tasks will be included: the test subjects will be allowed to comment on the query words and results etc.

Many different query modification strategies can be used to generate queries from the suggested terms – studying the different strategies is one of the goals of the study. The simulated search sessions make it possible to account for the users' evolving needs and relevance assessments by allowing the relevance criteria to evolve from liberal in the beginning of the session (evaluation of 1st generation queries) towards stricter, depending on the quality of results from each query round. Also the users' growing frustration potentially leading to abandoning the task can be modelled within this evaluation setting (also depending on the quality of results and cost of the query modification). The cost of the different actions (query formulation and modification and reading/assessing text snippets before relevance decision) are accounted for in the simulation of user interaction, e.g., the low quality of OCR text may lead to higher cost of reading and making sense of the results. Also, decision time related to assessing a document useful/useless varies depending on the task stage (exploration) and the grade of relevancy of the documents. Some assumptions made when setting up the simulation are summarized in Table 3.1.

Feature	**Value**	**Relation to evaluation**
Collection/language	Historical	Identifying good query terms difficult. Number of query modifications/time needed before acceptable result reached Added time and query rounds in the simulation.
Collection/quality	Mediocre: frequent OCR errors	Low OCR quality should be reflected in the cost estimations of the sessions as it makes identifying good query terms and making relevance assessments more difficult. Session length, user frustration.
Primary actor/domain expertise	competent	Affects relevance criteria, affects ability to recognize good query words and search strategy – should affect the costs in the simulation and the query modification strategies tested
Task type	Work task / academic: learning and writing	Affects search goal, relevance criteria and user motivation. What is enough? How much and how relevant information will make the primary actor satisfied?
Task stage	Exploration	Affects relevance: liberal criteria in early task stages, relevance assessment is difficult, when the goal and the relevance criteria is still unfocused

Goal	"enough" documents, low cost	Number of relevant documents seen by the user during the session, time, number of query modifications
Type of goal	Undirected, informational	Topicality, specificity and scope of documents
Urgency	Low	Max session time is rather high: results are not needed urgently
Motivation	High	Max session time: long sessions are possible, user tolerates some frustration
Result presentation	List of snippets of a few top ranked documents (OCR text)	Decision time for relevance

Table 3.1. Some assumptions and their relation to evaluation.

The results are measured using the GOMS (goals, operations, methods and selections) method (e.g. Smucker 2009) that can be used to find the sequence of actions (operations) that allows the user to achieve the user's goal in the shortest amount of time. In other words, GOMS allows including both what the user wants to find and the cost of the interaction in the evaluation. For example, it can be measured which sequence of query modifications would allow finding 10 relevant documents in shortest amount of time, or how many relevant documents can be found within a maximum time limit. Thus GOMS can be used to combine retrieval quality, a user model, and the hypothetical user interface in evaluation and to make a prediction concerning user performance (Smucker 2009).

4 Similarities and differences between use cases

So far, the use cases arising from the PROMISE use case domains seem to be very similar in many aspects. They describe the information access of professionals working on their work tasks, with high motivation and domain knowledge, working under good conditions with adequate devices and low constraints for input and result examination. Only one of the "additional" use cases, that of people search, describes information access in a non-work task related (leisurely) context. Mainly informational search goals are considered, with the one exception of a resource search for cultural heritage (downloading an image), even though the goals have quite different scopes: from finding one or few images in the medical use case to finding everything considering a topic in the intellectual property use case.

In most of the use cases, the time constraints are low. Even when these constraints are supposed to be moderate or high (the IP use case), the pressure from this limitation seems low as the task is a routine task performed 1-2 times each day. All the use cases describe search tasks with potentially long search sessions and many query modifications. Patent retrieval is the core work task of patent engineers in the intellectual property domain. Medical doctors regularly search for diagnosing support related to their work tasks. History

teachers sometimes search for images for illustrations of lectures. The main source of variation is the repository: the repositories discussed range from the whole web to well-structured collections with records consisting of text, images, and metadata. In all the presented repositories the primary object seems to be text anyhow – documents or metadata, even in those cases where images are primary input or important part of the result.

All of this is symptomatic of the classical ad hoc information access evaluations, where the primary actor is typically considered to be motivated, well-spoken and active searcher with clear, topical information need and high motivation and where the main variation between experiments is related to the text collections used. As a consequence of the similarities in the use cases, the use case framework is mainly tested from a limited perspective: for its suitability for description of professional information access use cases. This also results in limited creativity in developing new evaluation approaches, as the use cases can mainly be fitted into the typical Cranfield style laboratory evaluation settings. This is not really a limitation of the use case domains though: non-professional information needs and search tasks can be indentified for all domains.

Some differences (other than repository related) are described in the use cases. The use cases describe different phases of the search tasks and the primary actors have varying domain and search expertise. These are all features that affect the primary actor's ability to formulate queries and to make relevance assessments. Thus these differences should be reflected in the evaluations based on these use cases. Also, the people search use case describes a leisurely information access situation, where the primary actor is searching for information just for the sake of knowing, as opposed to searching to support some work task or problem solving. How this affects the primary actors success criteria and motivation should be considered in experimental design.

To summarize, one could say that we have so far only covered a small part of the use case space. The use cases discussed in this deliverable are quite similar in many respects. Various features and many types of use cases have not yet been considered. There are several compelling reasons to explore the space of possible use cases more thoroughly. First, in the near future, we can foresee new developments in both entertainment and information systems due to foreseeable advances in home digitalisation, in mobile and ubiquitous computing and in social informatics. This requires new thinking about human interaction with the internet of things. Second, extending research and development in information systems to new areas, we can see that there are numerous challenges ahead, related to evaluation and assessment of usefulness of systems. Cranfield style experiments may not always be a comfortable fit for evaluation, and researchers will have to innovate evaluation practice. Third, by working out for individual use cases how each feature value relates to choices in evaluation setup we may hope to learn recommendations for evaluation based on individual feature values. But if we are to learn how different feature values interact to inform evaluation design we would need to study a large body of well specified use cases indeed.

While we cannot expect to describe a large body of use cases without significant uptake of our use case framework outside and after the PROMISE project, we can list the main directions in which we first would like to explore the use case space. In future use cases professional tasks in non-academic or educational settings should be considered, as well as searching under sub-optimal conditions such as high time constraints, poor means for query input and/or result output (professional and leisurely tasks). More attention should be directed towards use cases with vague or non-topical information needs, use cases where the goal is to interact with a resource, being entertained or to navigate somewhere. Lower profile tasks, such as everyday and leisurely search tasks with low motivation or low task urgency should be considered. This might bring up many evaluation issues that cannot be solved in standard Cranfield style experiments, thus calling for extensions of the model or new approaches to evaluation. Thus, the next round of use cases from the PROMISE use case domains should probably concentrate on non-professional tasks. It also seems advisable to continue working on use cases from different domains, so that more variation can be covered.

Describing more use cases allows us to explore the boundaries of what we can describe with the use case framework we developed in Section 2, thereby in a sense validating the use case framework. Another way to validate the use case framework is to validate use cases described in it. The validation of use cases is an important goal in itself. The principal way in which we aim to do this is to describe use cases representing needs of actual users, and for which services are offered by actual stakeholders. By then interviewing these end users and stakeholders we validate use cases and the framework in the sense that we make sure the use cases accurately reflect users, their needs, and foreseen usage of the systems under scrutiny. This validation of the use cases is the subject of the next chapter.

One alternative form of evaluation that we are also developing in the PROMISE project is black-box evaluation, which is described in Section 6. Until we wrote this deliverable, it was an independent effort, inspired by earlier work done by partners in our network of excellence (Braschler et al, 200x). In Section 6 we explain this evaluation approach, and discuss how it can be adapted to, and informed by use cases described in the use case framework.

5 Involving stakeholders for the validation of use cases

The validation of the use cases seeks to determine whether they cover the requirements of the envisioned users of the targeted information access systems and whether they provide realistic descriptions of these systems' usage and behaviour. This is an iterative process; user requirements and system usage and behaviour have already informed the initial specification of use cases (see deliverable D2.1) and thus the development of the use case framework (described in Section 2) and the refined specification of use cases (presented in Section 3). The next step is to further validate this latest refinement with the goal to inform the final specification of use cases and evaluation tasks and also provide some further

feedback on the use case framework. The results of this validation will be reported in D2.4; this section describes the process to be followed to this end and discusses the most important issues to be considered at this stage.

The realism, accuracy, and coverage of the use cases will be further validated by interviewing envisioned users of the targeted information access systems and involved stakeholders. In particular, we intend to provide the interviewees with the use case forms (see Section 3) and ask them to validate our view of their use cases by filling in a structured questionnaire of a series of open-ended questions common to all use cases. This allows us to gain further insights into each use case through its validation, while it also enables us to make comparisons across the different use cases. Here, we describe the most important validation issues addressed by these questionnaires and the users/stakeholders to be targeted for each use case.

Once the purpose of the interview has been explained to the interviewees, they will be asked to go through the use case forms and then they will be asked to validate them by addressing the following issues:

1. **Use case description**: The description of the information retrieval situation will be validated in terms of its realism (does it reflect an existing situation?), accuracy (does it accurately describe that situation?), and coverage (does it cover all important aspects of this situation or have simplifications been made?). Further feedback on the importance, frequency, and prevalence of this situation for users/stakeholders will also be sought.
2. **System features**: The description of the system features will be validated in terms of its realism (does it correspond to information access systems they use?), accuracy (does it accurately describe such systems?), and coverage (does it cover all important aspects of such systems or have simplifications been made?).
3. **User features:** The description of the user features will be validated in terms of its realism (does it reflect themselves (for users) or their clients (for stakeholders)?), accuracy (does it accurately describe them?), and coverage (does it cover all their important features or have simplifications been made?).
4. **Session features:** The description of the session features will be validated in terms of its realism (does it reflect their goals when interacting with such systems and their interaction patterns with such systems?), accuracy (are these accurately described?), and coverage (does it cover all their requirements or have simplifications been made? What would be the ideal system for their needs? What would be the desired functionalities, input possibilities, and result formats?).
5. **Overall**: Finally, they will be asked to provided their overall opinion on the use case so as to gauge whether we have done a right job so far.

These questions are expected to help find out whether all user requirements have been covered by the refined specification and inform the final specification of the use cases and evaluation tasks.

The validation of the realism and accuracy of the existing use cases will in part also function as validation of the use case framework. As mentioned previously in Section 4, the use cases currently only cover a limited area in the use case space: the use cases are quite similar with respect to many features. Therefore, we aim to further validate the use case framework through using it for the description of markedly different information access use cases. These use cases can be identified through discussions and interviews with stakeholders from within the PROMISE use case domains and from outside and later validated through interviews as presented above. We are making an ongoing effort to locate and contact new stakeholders with use cases that have not been previously addressed. We strive to identify non-professional or otherwise varied use cases within the domains (such as the health related information access of the general public). Outside the domains, the focus should be on identifying as varied use cases as possible. Thus work on additional use cases is crucial for the validation of the use case framework.

We now list some users/stakeholders identified as the most appropriate for performing this validation. For the “Visual clinical decision support” us case: clinical practitioners, with a particular focus on radiologists. Further validation will be also be performed by comparing and contrasting the main points of the use case with the main findings of surveys conducted as part of the activities of the Khresmoi project (http://www.khresmoi.eu/) on the image search behavior of radiologists. For the “Search for innovation” use case: Patent searchers at patent offices or in private practice. The findings will be compared with a survey done in the IMPEx project on patent image search behaviour. (http://www.joanneum.at/?id=3922&L=0). For the “Search for Historical News” use case we intend to work with the National Library of Finland or the Royal Library in Sweden. An additional use case domain we are considering is “Entertainment for kids”, for which SVT bolibompa Web is a relevant stakeholder.

6 Black-box evaluation of information access applications

The main contributions in this deliverable are the use case framework described in Section 2, the specification of use cases in Section 3, and the specification of evaluation experiments evaluating part of the functionality described in the corresponding use cases, also in Section 3. Independent from the development of the use case framework, another evaluation effort has been started inside the PROMISE project: Black-box evaluation. Black-box evaluation of information access applications focuses on evaluating complete systems as opposed to evaluation only some components, e.g. a ranking engine. Evaluation is done from the user perspective, and aspects ranging from the quality of the collection being searched to user happiness are taken into account. In this section we explain and report on the status of the black-box evaluation effort in PROMISE. We highlight the many interesting relations between the use case framework and the black-box evaluation effort throughout the text and in a separate section at the end of this chapter.

6.1 A model of information access applications

In this section we describe the scope of the black box evaluation effort in terms of the kinds of information access applications we aim to evaluate. Concentrating on evaluation of complete information access applications, instead of specific search engine components, emphasizes the importance of high quality information access *application*s to enterprise (or industrial) information providers. The worth of individual components in a system should be assessed based on their effect on the end result, and the evaluations must incorporate the understanding that optimal performance of a component is not always necessary in face of other demands on the application. Therefore, testing and evaluation must be done not only on system components separately but also on a complete information access application, including the system proper, data and various configuration parameters of value for the service provided by the application to its customers. The idea of the black-box evaluation approach is to provide a generally usable methodology for information access application evaluation of operational systems and live systems as well as mirrored test environments. The approach of IR application evaluation is based on 3 premises:

1. *Evaluation is performed on a black box or minimally invasive*
2. *Evaluation is performed on operational applications*
3. *Evaluation is performed in an enterprise IR context*

In Figure 5.1, we present a model for the kind of information access applications we aim to evaluate.

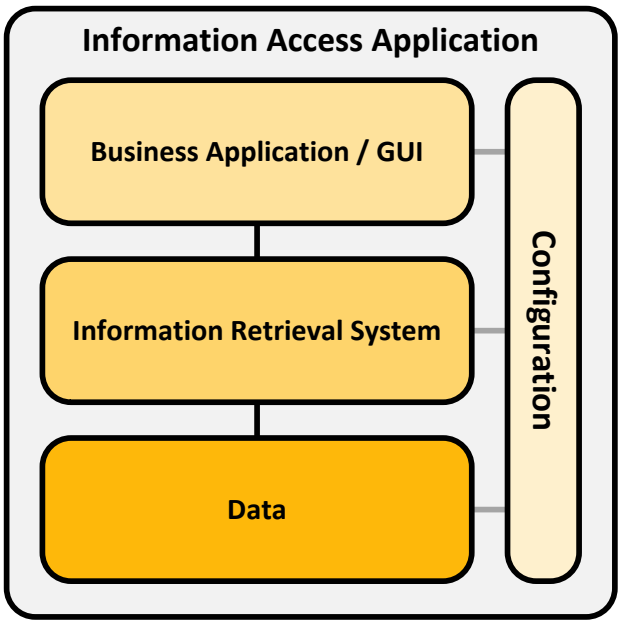

Figure 5.1: Information Access Application Model

The *business application* layer is made up of the user interface and associated business logic. Through this layer, user input and results presentation is handled. It also provides users with the means to interact with the IR system. Business logic may for example include facilities such as search sessions, search constraints or input validity checks. These processes are reflected in the business application layer and are a central part of the evaluation.

The *information retrieval system* layer represents the IR systems as understood in *Cranfield style* [Voorhees 2002] evaluations. The system layer is concerned with matching queries received from the business application layer to documents in the data layer. Thirdly, the *data* layer contains the search index and associated data interfacing and transforming functionality.

Parallel to these layers, the application's *configuration* represents operational parameters. The model thus acknowledges the importance of correct parameterization of applications according to the underlying business processes.

Users are not incorporated in this model of information access applications. We model user preferences later when we determine which aspects of applications we evaluate and how much influence each aspect will have on overall evaluation. More on that below.

In the next section, we discuss some quite general requirements for evaluation metrics for information access applications, and comment on the kinds of research questions that can be answered with such metrics. After that, we will describe the evaluation metrics we have in mind in the black-box evaluation setup.

6.2 Requirements for evaluation metrics

Now, what are then applicable metrics and measurements for doing IR application evaluation? Basically, the set of metrics and measures should be designed in such a way so it can find practically significant differences as opposed to merely significant differences between applications [Sanderson & Braschler 2009]. Also, an evaluation model and metrics in the context of information retrieval application evaluation need to provide absolute measures that are comparable across applications. A clear set of standards and measure thresholds serve as indicators of an estimate of user perception to evaluators. The aim is to provide corporate decision makers and others of operational information systems, with clear indicators of their applications' performance and to enable them to identify important issues as quickly and clearly as possible. Ideally, the inclusion of best practices would allow for specific recommendations of improvement.

Possible questions that could be addressed during an evaluation using this methodology are:

- How well does an application A perform by itself?
- Is application A better than application B?
- Is a new version of application A better than the old version of application A (small changes)?
- How does the performance of application A change over time?

In the black-box evaluation setup, we aim to define metrics for various aspects. In other words, we use various criteria for evaluating applications. In the next section we explain these criteria and their organization.

6.3 A hierarchy of evaluation criteria

The goal of black-box evaluation is to estimate the user perception of entire applications. Therefore, evaluation criteria need to cover the whole range factors from ergonomics on the user interface level to e.g. meta-data quality on the data level. We organize criteria in categories. These categories are top level aspects of applications which are evaluated by the weighing the criteria grouped under them. Preliminarily, we suggest using the following categories: Index, Matching, UI and Search Results. Then, for each criterion we define an evaluation metric in the form of a set of 'tests'. These tests are described in a protocol with the help of which assessors can evaluate applications. Eech test results in a numerical score, and test scores can again be weighed to arrive at a score for a criterion. We work this out in an example later in this chapter. By performing a large number of tests, the tests themselves can be kept fairly simple.

It is important to understand that the criteria, their weight, their associated tests and the weights of these tests, are all based on assumptions about user preferences. For each use case domain (such as the cultural heritage or intellectual property domain), the criterion hierarchy can be instantiated prior to evaluation to only include applicable criteria and tests. It is even possible to cater to individual use cases rather than use case domains. Criteria may be validated against use cases or use case domains by their associated assumptions.

Another way to say the same thing is that we model user satisfaction implicitly in black-box evaluation with the set of criteria, test, and their relative importance. Our test protocol is informed by the assumed user context and explicit and implicit knowledge a typical user would have in that situation. Since use cases according to the previously formulated framework in Section 2 of this deliverale supply this context, they are an integral asset for test protocol creation. The use case framework supports and informs the black-box evaluation methodology by providing structured information about use cases. This facilitates criteria/test elaboration as well as applicability assessments for previously known criteria. The use case framework describes, in detail, different user features. We will see later in this chapter how features such as role, types of expertise, language and demographic variables may be taken into account in a black box evaluation setup by selecting and weighing the criteria.

Furthermore, one can also extract tests from implicit knowledge and preferences which are common among some or all use case domains, that is, that some implicit knowledge and preferences are independent of the user or use case domain context, as we may for example assume that Google has set a very solid standard in terms of user interface for (text-based) search.

6.4 An abstract example of a hierarchy of criteria

In cases wherein multiple applications are being evaluated the evaluation structure can be visualized as a *grid*. In Table 6.1 below we show an example of this. Note that 'Area' denotes the 'Category' of the criterion. It has no weight, since the top level categories are not combined into a single evaluation metric. In this grid, we can see that the leftmost column shows the category (area) and its corresponding criteria with associated tests to be performed. The columns to the right of that column show the range of weights for each criteria and test. Here two columns to the right display the assigned weights for each application evaluated. The grid is an example for an entire evaluation campaign where multiple applications are evaluated.

Area/Criterion/Test	Weight	Score Application 1	Score Application …	Score Application n
Area	-	(0.5 + 0.67 + 1) / (1 + 1 + 1)	…	(0 + 0.33 + 1) / (1 + 1 + 1)
-Criterion1	(0.5 + 0.5)	0.5	…	0
--Test1	0.5	1	…	0
--Test2	0.5	0	…	0
-Criterion2	(0.33 + 0.33 + 0.34)	0.67	…	0.33
--Test1	0.33	1	…	0
--Test2	0.33	0	…	1
--Test3	0.34	1	…	0
-Criterion3Test	1	1	…	1

Table 6.1 An abstract example of an evaluation grid.

For the special case of only one application being evaluated (e.g. if a company uses the methodology for themselves), the grid is simplified to the form shown in Table 6.2:

Area/Criterion/Test	Weight	Score
Area	-	(0.5 + 0.67 + 1) / (1 + 1 + 1)
-Criterion1	(0.5 + 0.5)	0.5
--Test1	0.5	1
--Test2	0.5	0
-Criterion2	(0.33 + 0.33 + 0.34)	0.67
--Test1	0.33	1
--Test2	0.33	0
--Test3	0.34	1
-Criterion3Test	1	1

Table 6.2 An abstract example of an evaluation of just one system.

Tests are usually weighted the same within a criterion and their results are summed up in the criterion value. This leads to the weights of the associated tests being fractions of the weight of their associated criterion. Scores for tests range from 0 to 1. For example, a binary feature test would yield 0 for “not implemented” and 1 for “implemented”. In a test based on retrieving a set of 5 documents, each document successfully retrieved would increment the score with 0.2. In the next section we show an example of a criterion.

6.5 An example of an evaluation criterion

For each criterion in our hierarchy of evaluation criteria, we will list a category, an underlying assumption, ‘irregularities’, ‘root causes’ and a set of tests (testable features). Below we show an example criterion, the criterion ‘Completeness’ within the Index category.

Name:

Criterion Completeness

Category:

Index

Assumption:

Users expect to potentially find all documents that can be publicly accessed in any way on the site when using the search facility.

Irregularity:

Publicly accessible documents (known through browsing or obtaining a direct link) cannot be found using the search facility.

Root cause

The index is incomplete – documents/sets of documents are missing

The index is incomplete – the index is out of date (→ Freshness)

The index is incomplete – documents of certain types are missing (→ Format support)

Tests

1. Tester locates three documents that match the following criteria:
   a. at least 7 clicks to locate document
   b. document is at least 5 levels from root (as determined by URL)
   c. URL is at least 100 characters long
2. If no documents matching criteria 1) are found → abort
3. Tester extracts a characteristic phrase from the document
4. Tester searches for the document
5. Score: number of documents that can be located in the top 10 search results (0, 1, 2, 3)

Finally, we would like to describe a hypothetical example of using the guerrilla evaluation methodology in relation to the use case framework. For the PROMISE use case domain of Cultural Heritage (CH), an excerpt of the hypothethical grid is shown in Table 6.3.

The weights should be based on the use case domain leaders' sense of importance of a given criterion and are only for illustration in this example. In the next section we show a result obtained via discussion among 'use case domain owners' in our project: A list of criteria for evaluation of information access applications in all domains, together with a level of importance for each criterion.

Area/Criterion/Test	Weight	Score CH Application 1	Score CH Application 2
Index	-	**2.8**	**2.2**
- Duplicate Documents	1.5	**1**	**0.75**
-- Redundancy Test	0.75	0.33	0.66
-- Version Test	0.75	1	0.34
- Meta Data	2	**1.3**	**1.2**
-- Completeness Test	1	0.9	0.4
-- Correctness Test	1	0.4	0.8
- Freshness	0.5	**0.5**	**0.25**
-- News Test	0.5	1	0.5

Table 6.3 An excerpt of a hypothetical evaluation grid in the cultural heritage domain.

6.6 Criteria importance in three use case domains

Here, a preliminary assessment of importance of the identified criteria to the PROMISE use case domains is discussed. Making a round through all use case domains, the applicability and importance of the criteria were determined during a WP2 meeting on information retrieval application evaluation. Table 6.4 shows the discussed importance per criterion and use case domain, using simple indicators. In future work, when we will actually evaluate systems, this list of criteria may be tailored further to specific use cases, rather than use case domains. Note that in the table below, the criteria are not grouped yet under the four

main categories. Nor are tests specificied explicitly for each criterion. The main purpose of the table is to illustrate that the importance of criteria varies considerably over different use case domains.

6.7 On the relation with the use case framework

We have already noted that the structured contextual information provided in use case descriptions done in the use case framework developed in Section 2 can inform the selection and weighing of criteria in the black-box evaluation setup. Since the list of criteria used in the black-box evaluation effort is still under development, it can even inspire the definition of new criteria and tests. There are other differences and similarities of interest between the use case framework and the black-box evaluation effort, however.

Table 6.4 Preliminary list of criteria and their importance in the three main PROMISE use case domains.

Legend:

-- not important at all
- unimportant
0 neutral
+ important
++ very important

Criterion	Search for Innovation	Medical Image Retrieval	Cultural Heritage
Index completeness	0	+	+
Index freshness	0	+	-
Binary document handling	+	+	+
Separation of actual content and representations	-	-	0
Special character handling	+	+	0
Synonyms, domain specific terminology	+	++	+
Duplicate documents *	++	+	+
Meta data quality	+	0	++
Tokenization	+	+	+
Enrichment	++	+	++
Stability	++	-	-
Phrasal queries	++	++	+
Query syntax	+	+	-
Over-/underspecified queries	0	0	++
Feedback	+	+	++
Multimedia	0	+	++
CLIR	+	0	++
Facets	++	+	+
Search in fields	+	0	-
Performance / responsiveness	+	+	+

Criterion	Search for Innovation	Medical Image Retrieval	Cultural Heritage
User guidance	+	+	+
Browsing	+	+	++
Personalization **	-	0	+
Social media	+	+	+
Error handling	+	+	+
Entertainment / fun	-	-	++
Localization	+	+	+
Result list export / import	++	+	+
Sorting of results	+	+	+
Justification of search results	+	+	+
“Monitoring”	+	+	+
Override silly system actions (user control)	+	+	+
Related content	+	+	++
Context information	+	+	++
Navigational aids	+	+	+
Navigational queries	0	0	-
Factual queries	0	0	-
Informational queries	+	+	+
Known item retrieval	+	-	++
Diversity	0	0	+
“Linguality”	+	0	+
(Geo-)Location	--	-	+

Table 6.4 Preliminary list of criteria and their importance in the three main PROMISE use case domains.

* There is a difference between duplicate documents and versions of the same document. Redundant identical documents are unwanted while duplicate (or very similar) information is useful.

** The criterion “personalization” contains several different aspects: Look and feel, personal profile and treatment in search, control of application. Different types of users appreciate such features differently.

The black-box evaluation effort aims at defining a relatively stable and exhaustive set of criteria by which information access applications may be evaluated. The use case framework is meant to be a relatively stable hierarchy of features by which use cases involving information access applications may be characterized. The main difference between the two efforts is that the use case framework is not an evaluation experiment, but the black box evaluation is. By capturing the context in which systems are used, the use case framework can inform which aspects should be evaluated. The black-box evaluation setup is an evaluation framework that has many parameters which can be adjusted to cater to a wide variety of use cases. These parameters are the weights of criteria.

There is conceptual overlap between use case features and black-box evaluation criteria. This facilitates adjusting the black-box evaluation framework in a straightforward way by weighing the importance of criteria. There is overlap between the two approaches in the

way they model repositories, functionality of potential systems under scrutiny, and foreseen usage. We give some examples of each. In the 'Repository' section of the 'Secondary actors, the features 'Indexing timeliness' and 'media' are strongly related to the criteria 'Index freshness' and 'Multimedia' in Table 6.4. Export functionality of search result pages is also foreseen in both the use case framework (Session features -> Elements of interaction pattern -> Export) and black-box evaluation (Result list export / import). Examples of foreseen usage are the kind of queries expected: navigational, informational, or known item in the list of criteria in Table 6.4, whereas in the use case framework the overlapping taxonomy of Rose & Levinson (2004) is used.

In their future development in the PROMISE project, the use case framework and black-box evaluation effort can benefit from each other. By comparing the features in the use case framework with the criteria in the black box evaluation effort both the list of features and the list of criteria may improve. For the use case framework, it is important to describe more use cases in it, and validate more evaluation experiments with it. A next step for black box evaluation is to perform evaluation experiments with it. With the use case framework in hand it can be studied if these evaluation experiments will reflect user satisfaction. For the validation of use cases described in the framework it is important to work with real service providers (stakeholders) and real users. For black-box evaluation evaluating operational systems of stakeholders is equally important.

7 Discussion

We have motivated and developed a use case framework for describing use cases involving information access applications, building on the groundwork done in D2.1 [Karlgren et al, 2011]. We adopted an iterative approach to describe use cases and develop the use case framework. The framework in Section 1 and the use cases in section 2 reflect the current status of this work.

While the use case framework is intended to remain open for alterations and additions, we feel that it has reached a somewhat stable form. Nevertheless, in Sections 4 and 5 we have noted that there is a need to describe more and more varied use cases to test the generality and applicability of the framework. These use cases might still result in notable changes in the framework. Therefore, it is urgent for us to contact more stakeholders, and to describe more use cases in the framework.

With the structure and the features of the framework somewhat stabilised, we can now move the focus of the work to mapping the use case features to evaluations decision and benchmarking mechanisms. Looking at the use cases and evaluation tasks described in Section 3, we see that more thorough discussions of each of the use case features from the perspective of evaluation are needed. Table 2.11 in Section 2.5 may be a starting point in this effort.

A first step in this direction will be formulating hypotheses concerning user preferences and the most central success criteria for each use case: each of the use case domains could consider for each feature in their use case(s) how it affects user interaction with the system and how it should be evaluated. When each of the features has been considered in isolation, the interaction effects of the different features need to be regarded: the different features interact in various ways and a single feature can motivate many different evaluation approaches depending on the rest of the use case. Creating consistent experiments and evaluation tasks based on features pointing to various directions requires compromising and formulating hypotheses concerning the preferences and most important success criteria of the end users. Just getting here is a valuable goal in itself: When use cases with explicit hypotheses concerning the system usage and evaluation criteria are 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 build and design their systems according to results given by the use case, if they find it conforms to the behaviour they can observe in their customers and clients (Karlgren et al, 2011). Practical experience from applying the framework to real world tasks and judicious modelling of the generalities and defining characteristics of the tasks is needed to understand the complex interactions between the different features.

A second step, which is a long term goal of our work, is to enable the future validation of hypotheses concerning user preferences discussed above. Our belief is that hypotheses of user preference, task model, and interaction in sessions should at some point during a system development process be validated in the sense that it should be established that they hold for real end users in the wild. This is a tall order and out of scope for this project—it is really a task for interaction specialists. It is not necessarily a task necessary to solve while building an information access system, but it is necessary before moving it to practical application. We noted in the introduction that user studies may be the most powerful way to validate hypotheses about user preferences, but they have their own problems: they are expensive and have to deal with large variance in individual user behaviour. They are best performed by user study professionals, not information system builders. Other means to start the validation of hypothesized user preference are literature reviews, introspection, stakeholder interviews and end user interviews. In addition interactive tasks in an evaluation campaign are an interesting possibility.

A third step, also a long term goal, is to obtain benchmarking and best practice recommendations based on use case features, the former for developers, the latter for information technology professionals in need of support when purchasing components for a system. As noted in the introduction, their quantity and quality will increase as more use cases are described with the use case framework. Then one can start looking for similarities and differences in use cases: do similar features result in similar benchmarks? This work is related to the work done in Task 2.5 (Best Practices in Multilingual and Multimedia

Information Access), where the aim is to formulate best practices based on the output of evaluation tasks and activities.

A next challenge for the short term in WP2 is to find a very condensed form for the specification of test collection based evaluation tasks (as opposed to use cases) that will enable organizers of these tasks to demonstrate how choices in evaluation setup are motivated by the underlying use cases. If we want to motivate organizers from outside the PROMISE project to do this, we will have to find a minimal set of use case features, aspects of evaluation tasks and relations between them that is still useful. Our aim for CLEF 2012 is to ask all lab owners for a one page specification of their evaluation task, with the underlying use case in mind.

Meanwhile, we believe that the use case framework is already useful, because it forces researchers and system developers to think about all aspects that play a role in the foreseen use of systems. Our short term goals focus on consolidating and refining the framework through ongoing discussion amongst PROMISE partners and with stakeholders, on describing new and varied use cases beside the three main PROMISE use case domains, and on doing more analysis on evaluation tasks targeting described use cases: specifically how do use case features relate to evaluation decisions?

The black box evaluation effort has the ambitious aim to be applicable to all use case domains, and estimate user satisfaction both in an absolute sense, and in a relative sense (as user preference for one system or another). Already there is a list of proposed evaluation criteria, and a preliminary weighting for each of the three main PROMISE use case domains. An interesting possibility is to evaluate a set of systems which are also evaluated in benchmarking style. This would give us the opportunity to compare the two evaluation methodologies, giving rise to many research questions, such as: to what extent is the quality of a ranker as measured through e.g. MAP reflected in the perceived quality of the category 'Search Results' in a black box evaluation? The obvious first priority of the black box evaluation effort it is however to do a number of evaluation experiments on operational information access applications.

References

Agirre et al., 2009 — Agirre, E., Di Nunzio, G., Ferro, N., Mandl, T., & Peters, C. (2009). *CLEF 2008: Ad Hoc Track Overview*. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. Jones, M. Kurimo, T. Mandl, A. Peñas & V. Petras (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access. CLEF 2008 (Vol. LNCS 5706, pp. 15-37). Berlin / Heidelberg: Springer.

Bailey et al, 2007 — Bailey, P., Craswell, N., de Vries, A. P., & Soboroff, I. (2008). *Overview of the TREC 2007 Enterprise Track*. In The Sixteenth Text Retrieval Conference Proceedings (TREC 2007). NIST. Special Publication.

Balog et al, 2007 — Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., & van den Bosch, A. (2007). *Broad expertise retrieval in sparse data environments*. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIRconference on Research and development in information retrieval, (pp. 551–558). New York, NY, USA: ACM Press.

Balog et al, 2008 — Balog, K. *People Search in the Enterprise*, Ph.D. thesis, University of Amsterdam, 2008.

Balog et al, 2009 — Balog, K., Soboroff, I., Thomas, P., Craswell, N., de Vries, A. P., & Bailey, P. (2009). *Overview of the TREC 2008 Enterprise Track*. In The Seventeenth Text Retrieval Conference Proceedings (TREC 2008). NIST. Special Publication.

Berendsen et al, 2011 — Richard Berendsen (UvA), Giorgio Maria Di Nunzio (UNIPD), Maria Gäde (UBER), Jussi Karlgren (SICS), Mihai Lupu (IRF), Stefan Rietberger (ZHAW), Juliane Stiller (UBER), *Deliverable 4.1 - First Report on Alternative Evaluation Methodology,* 2011.
http://promise-noe.eu/deliverables.

Berendsen et al, 2011 — Berendsen R., Kovachev B., Meij E., de Rijke M., Weerkamp W., *Classifying Queries Submitted to a Vertical Search Engine*, Web Science 2011, Koblenz, ACM, June, 2011.

Braschler et al. 2009 — Braschler, M.; Heuwing, B.; Mandl, T.; Womser-Hacker, C.; Herget, J.; Schäuble, P.; Stuker, J. (2009). *Evaluation der Suchfunktion deutscher Unternehmenswebsites*. In: Wissensorganisation 09: "Wissen – Wissenschaft – Organisation" 12. Tagung der Deutschen ISKO (International Society for Knowledge Organization) (19-21.10.2009 in Bonn)

Broder, 2002 — Andrei Broder. 2002. *A taxonomy of web search*. SIGIR Forum 36, 2 (September 2002), 3-10. DOI=10.1145/792550.792552 http://doi.acm.org/10.1145/792550.792552

Craswell et al, 2006 — Craswell, N., de Vries, A., & Soboroff, I. (2006). *Overview of the TREC-2005 Enterprise Track*. In The Fourteenth Text Retrieval Conference Proceedings (TREC 2005). NIST. Special Publication.

De Rijke et al, 2010 — de Rijke M., Balog K., Bogers T., van den Bosch A., "*On the Evaluation of Entity Profiles*", CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation, Padova, Springer, September, 2010

Ferro et al., 2010 — Ferro, N., & Peters, C. (2010). *CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks*. In C. Peters, Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., & G. Roda (Eds.), Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross-Language Evaluation Forum (CLEF 2009). Revised Selected Papers (LNCS 6241, pp. 13-35). Heidelberg, Germany: Springer.

Guldbæk Rasmussen et al. 2010 — Guldbæk Rasmussen, K.; Iversen R.; Petersen, G. (2010) EuropeanaConnect: M3.2.3 *Personas Catalogue*

Harman & Buckley, 2009 — Harman, D., & Buckley, C. (2009). *Overview of the reliable information access workshop*. Information Retrieval, 12, 615–641. 10.1007/s10791-009-9101-4. URL http://dx.doi.org/10.1007/s10791-009-9101-4

Karlgren et al, 2011 — Jussi Karlgren, Gunnar Eriksson, Madlen Frieseke, Maria Gäde, Preben Hansen, Anni Järvelin, Mihai Lupu, Henning Müller, Vivian Petras, Juliane Stiller, *Deliverable 2.1-- Initial specification of the evaluation tasks, 2011* http://www.promise-noe.eu/deliverables

Kazai et al., 2010 — Kazai, Gabriella; Koolen, Marijn; Doucet, Antoine; Landoni, *Monica: Overview of the INEX 2010 Book Track. At the Mercy of Crowdsourcing*. In: Shlomo, Geva; Kamps, Jaap; Schenkel, Ralf; Trotman, Andrew (eds.): INEX 2010 Workshop Pre-Proceedings. December 13-15, 2010. Huize Bergen, Vught, the Netherlands.

Kazai & Doucet, 2007 — Kazai, Gabriella; Doucet, Antoine: *Overview of the INEX 2007 Book Search Track* (Book Search '07). In: Focused Access to XML Documents. Lecture Notes in Computer Science, 2008, Volume 4862/2008, 148-161

Keskustalo et al, 2009 — Keskustalo,H. & Järvelin, K. & Pirkola, A. & Sharma, T. & Lykke Nielsen, M. (2009). *Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case of Extremely Short Queries*. AIRS 2009, the 5th Asia Information Retrieval Symposium, Sapporo, Japan, October 2009. Heidelberg: Springer, Lecture Notes in Computer Science vol. 5839, pp. 63-74.

Minerva Working Group 5, 2008 — Minerva Working Group 5 (Ed.). (2008). *Handbook on cultural web user interaction*. Retrieved August, 2011 from http://www.minervaeurope.org/publications/Handbookwebuserinteraction.pdf.

Rose & Levinson, 2004 — Daniel E. Rose and Danny Levinson. 2004. *Understanding user goals in web search*. In Proceedings of the 13th international conference on World Wide Web (WWW '04). ACM, New York, NY, USA, 13-19. DOI=10.1145/988672.988675 http://doi.acm.org/10.1145/988672.988675

Sanderson and Braschler 2009 — Sanderson, M.; Braschler, M. (2009) *Best Practices for Test Collection Creation and Information Retrieval System Evaluation*.

Sanderson, 2010 — Sanderson, M. *Test collection based evaluation of information retrieval systems,* 2010, Now Publishers

Smith & Kantor, 2008 — Smith, C., & Kantor, P. (2008). *User adaptation: good results from poor systems*. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, (pp. 147–154). ACM.

Turpin & Scholer, 2006 — Turpin, A., & Scholer, F. (2006). *User performance versus precision measures for simple search tasks.* In Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, (pp. 11–18). ACM

Weerkamp et al, 2011 — Weerkamp W., Kovachev B., Berendsen R., Meij E., Balog K., de Rijke M., *People Searching for People: Analysis of a People Search Engine Log*, 34th Annual International ACM SIGIR Conference (SIGIR 2011), Beijing, ACM, pp. 45–54, July, 2011.