PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 2.4
Use case inventory and final specification of the evaluation tasks

Version 1.1, March 2013

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:	2.4
Deliverable title:	Use Case Inventory and Final Specification of the Evaluation Tasks
Delivery date:	28/02/2013
Lead contractor for this deliverable	UGOT
Author(s):	Anni Järvelin, UGOT; Richard Berendsen, UvA; Gunnar Eriksson, UGOT, Preben Hansen, UGOT; Karin Friberg Heppin, UGOT; Jussi Karlgren, UGOT; Vivien Petras, UBER; Maria Gäde, UBER; Mihai Lupu, TUW; Florina Piroi, TUW; Alba Garcia Seco de Herrera, HES-SO, Stefan Rietberger, ZHAW, Martin Brachler, ZHAW.
Participant(s):	UGOT, UvA, HES-SO, TUW, UBER, ZHAW
Workpackage:	2
Workpackage title:	Stakeholders Involvement and Technology Transfer
Workpackage leader:	UGOT
Dissemination Level:	PU – Public
Version:	1.1
Keywords:	Information access evaluation, use cases, evaluation tasks, validation

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0	22/02/2013	Draft		
1.1	15/03/2013	Final		

Abstract

If we wish to see the research efforts in the field of information access to continue being relevant to commercial service providers, the scope of the research efforts in the field need to be broadened to better capture the mechanisms for information access systems' impact, take-up and success in the marketplace. We suggest that use cases offer a means of establishing the relevant success criteria for the systems and can thus guide the evaluation of information access systems. In this report, the final results of the work on use cases in PROMISE are reported: the validity of use case framework and use cases presented in deliverable D2.2 is tested, and the framework and the use cases are revised in accordance with the validation results. Finally, evaluation tasks are formulated based on each of the use cases and the connections between the features of the use cases and the evaluation tasks are discussed.

PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

Table of Contents

Executive Summary

Work package 2, Stakeholder Involvement and Technology Transfer, has the goal of increasing the stakeholders' interest in PROMISE evaluation activities and take-up of the evaluation results through involving the stakeholders in the formulation of the evaluation activities. Much of this work has been carried out through iteratively formulating and validating use cases for the three PROMISE use case domains of visual clinical decision support, unlocking culture, and prior art search, in continuous contact with the use case stakeholders. One of the major tasks in work package 2 has then been developing a framework that can guide the definition of use cases for the purposes of multimedia information access system evaluation: a framework for describing the constraints and demands related to the users and usage of various information access systems and for integrating these constraints to benchmarking mechanisms, to establish realistic and relevant success criteria for the systems, and to bring together benchmarking with system and service validation.

This deliverable reports the final results of the work on formulating use cases for information access evaluation. First, the validation of the use cases and the use case framework presented in deliverable D2.2 is discussed. While the validation results were overall positive, some issues for improvement were identified both for each use case domain, and for the use case framework. Consequently, a revised version of the framework is presented in the deliverable, followed by the final use cases for each use case domain. Even two novel use case domains related to reputation management and enterprise search are presented. The revised use case framework improves on several fronts: Most importantly, the framework is extended to provide tools for deriving evaluation tasks based on the use cases through mapping the use case features to experimental design decisions and benchmarking mechanisms. Better means for describing user-system interaction are provided, and the framework is made clearer and easier to use. The framework is re-organized into "background features", "interaction features", "system features" and "evaluation features" and checklists are provided for each set of features to make the description of use cases easier. Finally, evaluation tasks are specified for each of the five use cases using the new use case framework. These evaluation tasks show how the framework can support the formulation of broader and richer benchmarking experiments, where more focus is put on the end users' tasks and goals.

1 Introduction

Information access is no longer only a question of retrieving topical documents in a work-task related context. Document retrieval has become an embedded component in many systems which neither to their users nor their providers appear to be classic document retrieval systems: entertainment systems, communication platforms, time management systems, and the like. If we wish to see the research efforts published in the field to continue being relevant to commercial service providers, even the scope of the research efforts must be broadened to better capture the mechanisms for the systems' impact, take-up and success in the marketplace.

This is not unknown in the field. Many efforts in recent years have contributed to a richer understanding of users, their intentions, search sessions, and the evaluation thereof in formal, quantitative, or qualitative ways [e.g. Azzopardi 2011, Keskustalo 2009, Liu 2010, Smucker 2012]. However, this is not enough. What is still needed is a *framework* that can integrate the constraints and demands related to the users and usage of (the embedded information access components of) various systems, in various contexts and domains, and varied user communities, to evaluation mechanisms, and thus support richer and broader benchmarking, and bring together benchmarking with system and service validation. Developing such as framework has been one of the major goals in Work package 2. This report presents the final results of the work on developing a framework based on use cases and user centred design principles.

Use cases are a software development methodology, first developed by Ivar Jacobsson and colleagues [Jacobsson 1987, 1992] for capturing interaction-based functional requirements in software development, and further developed by others, e.g. [Cockburn 2002, Constantine 2006]. The requirements are captured from the user perspective by describing how a user interacts with a system to carry out a task, or to reach a goal. The focus is then on task modeling[1] or modeling one kind of use that a system can be put to, given a specific user role. In many cases, one user can use a system in several ways and for different purposes. Focusing on specific kind(s) of system usage in evaluation, instead of trying to cover all possible different interactions and goals in one experiment is then practical.

In work package 2, the goal has been to involve stakeholders in the formulation of the evaluation activities of PROMISE in order to formulate more realistic evaluation tasks and methodologies. This can increase the stakeholders' interest in and take-up of the evaluation results. The main vehicle of this work has been the iterative formulation of use cases on the three PROMISE use case domains of visual clinical decision support, unlocking culture, and prior art search. A great deal of work has been devoted to identifying the central features of the use cases, and to formulating the features into a framework that can guide the definition of use cases for the purposes of multimedia information access system evaluation. In the framework, observable patterns of human information access behavior are described through a selection of variables that can be linked to the features of experimental design and the system and interface features of the evaluated systems, as illustrated in Figure 1: The framework integrates the features affecting information access system usage, with the constraints presented by the system and interface design on one hand and experimental design on the other. This way the framework can indicate evaluation approaches for measuring the value of an information access system to its users given some real-world constraints of the system usage. It can describe to what kind of real-world information access system usages the results of a specific experiment can apply to.

[1] The use of "task" in use case contexts differs from the use of "work tasks" in information access literature: use cases focus on users' immediate tasks when interacting with systems, the task the user expects the system to support and not the broader work tasks that the users are engaged in.[1]

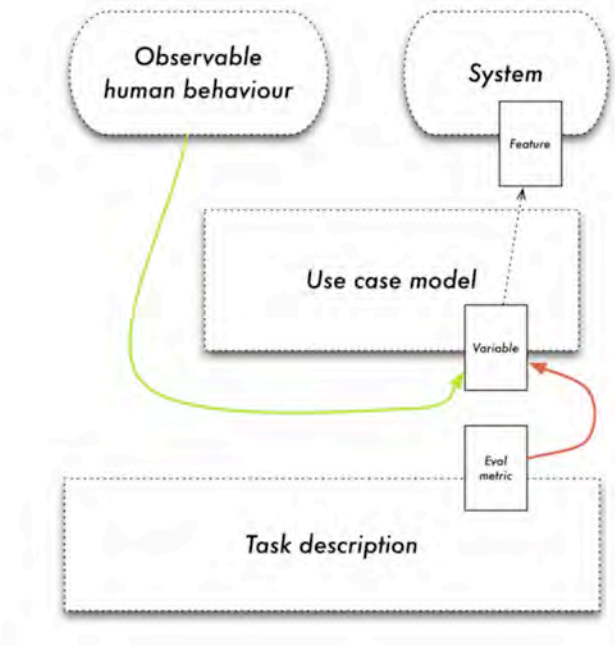


Figure 1. Relating human behavior to system and evaluation features.

Deliverables D2.1 and D2.2 described the first steps towards this goal: Deliverable D2.1 described the results of an initial requirements analysis for the focus domains and introduced the idea of a "use case model". Thereafter deliverable D2.2 developed an initial use case framework for supporting the specification of use cases and presented a set of use cases based on the framework.

This deliverable continues where deliverable D2.2 left off: The experiences from D2.2 suggested that the use cases and especially the use case framework needed another round of validation and revisions. A validation protocol based on structured interviews of the stakeholders and end-users of the evaluated systems was proposed in deliverable D2.2. The results of these validation interviews are reported in this deliverable. The validation results prompted a few modifications of the use cases and especially of the use case framework. Consequently, this deliverable presents a revised version of the framework, where especially the description of the user-system interaction has been given more attention and support. Further, checklists for defining the use case features, and better guidelines for following the use case methodology are presented in order to make the work of writing use cases easier.

The use case framework in deliverable D2.2 did not yet offer support for deriving evaluation tasks based on the use cases. In this deliverable the use case framework is extended to support mapping use case features to experimental design decisions and benchmarking mechanisms. A set of final use cases and specifications of evaluation tasks following the new use case framework is then presented. Two of these use cases and evaluation tasks are new to this deliverable: the enterprise search and reputation management use cases. The addition of the new, quite different use cases tests the suitability of the framework to describing information access evaluation tasks previously not considered.

2 Validation

The validation task T2.3 has sought to determine whether the use cases specified in tasks T2.1 and T2.2 cover the requirements of the stakeholders of the targeted information access systems and whether they provided realistic descriptions of these systems' usage and behavior. It has also sought to determine whether the use case framework developed supports formulation of valid information access use cases and evaluation tasks. The validation has been an iterative process: user requirements and system usage and behavior have already informed the initial specification of use cases (see deliverable D2.1) and thus the development of the use case framework (see deliverable D2.2) and the refined specification of use cases (see deliverable D2.2). The goal of this final round of validation was to further validate and, if necessary, improve the *realism, accuracy* and *coverage* of the use cases before the final specification of the use cases and evaluation tasks. It was also to provide further feedback on the use case framework.

As foreseen in deliverable D2.2, the final validation was carried out by interviewing a group of use case stakeholders, who had not been involved in the previous use case requirement analysis phase. Each interviewee was presented with a use case description (from D2.2) and then asked to evaluate it by answering a structured questionnaire.

The questionnaire contained 5 parts following the structure defined in deliverable D2.2: use case description, system features, user features, session features, and evaluation task. The first four parts discuss the use case and were divided into three subsections of questions concerning the realism, accuracy, and coverage of the use case section. The final part asked for the stakeholders' view of the usefulness of the evaluation tasks defined, based on the use cases: if the tasks target interesting issues and measure them in a reasonable way. All in all the survey contained 49 multiple choice questions and 2 open ended questions – one for general thoughts concerning the use case, and one for the evaluation task. Each use case had their own questionnaire form and collected their own data separately. An example of the complete questionnaire can be found here: https://docs.google.com/spreadsheet/viewform?formkey=dHNDeDV3elFzQnFvQnVTcklUU2l2QXc6MA#gid=0.

The results were analyzed for each individual use case (fed back to task T2.1), but also comparisons between the use cases were made to gain insight on how well the use case framework supports the formulation of use cases. The goal was to identify unnecessary or missing use case features, ambiguities or redundancies (fed back to task T2.2). Both the individual use cases and evaluation tasks, and the use case framework are updated in this deliverable based on the results (see sections 3 and 4). Finally, after updating the use case framework a stakeholder of the reputation management use case was interviewed to test the new use case framework.

Attracting stakeholders not previously involved in the development of the use cases proved difficult. Also, interviewing stakeholders not familiar with the use cases, PROMISE and CLEF required extensive explanations of many concepts before and during the interviews, making the interviews long and laborious. These factors limited the number of participants. In total, 14 stakeholders or potential end users participated in the validation: 13 filled in the questionnaire, and one was interviewed following the new use case framework structure.

The questionnaire was found problematic by some interviewees (and interviewers) for both the cultural heritage and the intellectual property use cases. Some questions and answer scales were found confusing and thus the interviewees were uncertain of how to answer. The interviewees very often disagreed in their answers, and the answers for a question concerning a single use case could range from very negative to very positive. Thus the results are inconclusive and not generalizable. However, they do give an idea of the general validity of the use cases and indicate issues that need further attention in the use cases and in the use case framework: A few problems and issues were identified for each use case. Some were common for all use cases, and therefore likely to depend on the use case framework or the way it has been applied. A detailed description of the results follows.

2.1 Medical domain: visual clinical decision support

There were four respondents for the visual clinical decision support for the medical diagnosis use case, who were either medical doctors or researchers in the medical imaging domain. They filled in the questionnaire on-line and did not report any problems with it.

2.1.1 Use Case Description

The use case description was very positively scored and considered to describe a realistic situation and a complete sequence of events without many simplifications. However, the possible variations of the flow of interactions and the points of interaction where they may occur are not adequately described (2/4 and 1/4 responses negative, respectively).

2.1.2 System Feature Description

The system feature description was also very positively rated. It was found to be a realistic and accurate system description, identifying correct secondary actors and system utilities. The only (weakly) negative answers considered the definition of system boundaries and coverage of all necessary system features, but even there only one respondent was critical, while others were very positive. None of the respondents found that there were simplifications made in the system description (even if the coverage was slightly criticized by one).

2.1.3 User Feature Description

The description of the user features was evenly positively scored. The respondents agree that correct users are described realistically and at an appropriate level of detail. The only (weak) negative answer was concerned with the simplifications made in description of user features: one of the four respondents indicated that simplifications have been made. This result does not concur with the results for the accuracy or coverage of the system and session feature descriptions.

2.1.4 Session Feature Description

Most problems were identified with the description of the session features. It was indicated that the system-user interaction was not accurately described and not at an appropriate level of detail. Also, some simplifications were identified in the description of the session features. Thus further information seems to be required concerning the interaction. However, this problem was only identified by one of the respondents, while the three others were generally very positive. The description of user goals was found very realistic.

2.1.5 Evaluation Task

Only one respondent keeps track of the evaluation task even if the problems, technologies, and user groups targeted by the evaluation are relevant according to all respondents. All the participants mean that the Case-based Retrieval Task is the most relevant. They are equally interested in the evaluation of mature and of new and experimental technologies. The participants also agree that the document collection contains realistic data and most of them understand how the ground truth is created for the test collection. Finally, they think that the measurement of clinical accuracy in the search is missing as well as the query times and index sizes.

2.1.6 Summary

The overall scores were positive although there are many disagreements in the answers. Due to the limited number of responses the results are inconclusive, but they do indicate that the interaction sequences, including their variations could be better described. One respondent also suggested inclusion of the administrative internal affairs of a hospital with the handling of a patient (e.g. referrals and legal implications) to the use case.

2.2 Intellectual property

Four experts (examiners, service providers and consultants) answered the final validation questionnaire. The use case for the intellectual property domain has been validated in a continuous dialogue with stakeholders since deliverable D2.2. In addition to the final validation questionnaire for which the results are described here, two additional surveys have been carried out: a large survey of 86 questions with a “major patent office”, partly overlapping with the final validation questionnaire; and a smaller interview with four patent examiners at the Greek patent office, containing the same four use case sections as the final validation questionnaire, but with open ended questions related to the realism, accuracy and coverage of the use case.

2.2.1 Use Case Description

The use case description was generally found to be realistic, accurate, and logical. One respondent however considered the described sequence of interaction to be incomplete. The points of interaction where variations may occur could have been better described, and the level of detail could have been better. Two respondents found that the use case makes simplifications with respect to the real task.

2.2.2 System Feature Description

The description of system features was again overall positively rated. Two respondents found some necessary aspects of systems missing from the description, but only one was slightly negative to the description’s overall correspondence with realistic systems, and only one clearly stated that simplifications have been made in the system description.

2.2.3 User Feature Description

Three of the respondents were overall positive to the user feature description, while one found that the described users maybe were not the correct, or realistic users (score: 4/10), that the description of the users and their context wasn’t accurate or on appropriate level, and did not cover all important user features (3/10).

2.2.4 Session Feature Description

The session features were also overall positively rated: All but one respondent gave positive responses to all questions. One respondent indicated that correct user goals had not been identified, that the user-system interaction was not correctly described, that the elements of the interaction pattern were not well defined, and that not all important session features were covered.

2.2.5 Evaluation Task

All the respondents were familiar with and kept track of the intellectual property evaluation tasks. Unsurprisingly then they found the problems, technologies and user groups targeted in them relevant. The experimental settings were found reasonable by most, though one respondent found especially the data, the ground truth creation and the way results are measured unrealistic. More user involvement in evaluation was wished for by one respondent.

2.2.6 Summary

The open ended question concerning the respondents’ overall opinion of the use case gives an indication of what the respondents finally considered to be the most important shortcomings of the use case: The use case was considered (by three respondents) to be somewhat generic and a simplification of actual prior art search, missing to indicate alternative (but important) interaction patterns (workflows, information sources, tools, search strategies). Therefore, it seems that more detailed description of the interaction sequences and clearer indication of the points where alternative interaction sequences may occur are needed. Also some detailed improvements were suggested by individual respondents: definitions of expertise and language skills could be improved, task frequency should be

reduced somewhat. These results were very much in line with the results from the two previous validation efforts: they were also very positive overall and identified the generic level of the description of interaction as the main problem leading to failure in capturing the specifics of any task.

2.3 Unlocking culture: the search for lecture material

In total, five participants were interviewed; four of them were directly involved in system development and one participant had a museum background. Participants were either interviewed face to face or via Skype. To begin with, the purpose of PROMISE, the use case framework, the use case and the CHiC lab were explained and potential questions clarified. During the interview participants were encouraged to ask further questions if they were not sure about the meaning of a question. On average each interview lasted around 60 min.

2.3.1 Use Case Description

In general, participants were satisfied with the phrasing and coverage of the use case description. One participant estimated the use case as totally realistic, while another stated that the use case does not reflect a realistic situation. It was stated that the use case is limited to a particular system type (in this case Europeana) and does not necessarily describe other CH systems. While everyone considered the sequence of interactions to be relatively complete, they all found that variations of the sequence and their possible occurrences were not described clearly enough. For example, additional discovery or exploring scenarios including saving of images or social media sharing are not included in the use case description. Furthermore, language or legal restrictions are not mentioned in the use case, but are considered important aspects within the CH domain.

2.3.2 System Feature Description

Within this part of the questionnaire a range of answers was observed. A more detailed description of possible interactions between the system and user is desired. In addition, important technical requirements for the use and reuse of the system and the data are missing. It is not stated to what extend external services like social media portals or web services could be integrated into the use case. Neither were system and interface design issues mentioned nor discussed with regard to interaction constrains and system boundaries.

2.3.3 User Feature Description

The user description was considered as complete, but not all participants regarded the identified user as realistic. While the user description itself was realistic, it was not considered to describe a typical user type of the particular system (Europeana). The use case should consider user groups that fit to the identified system under discussion.

2.3.4 Session Feature Description

Participants asked for a more detailed description of interaction patterns including boundaries and system specific constraints. Only one user could not identify the user goal (s).

2.3.5 Evaluation Task

The most relevant evaluation task is the Semantic Enrichment task, followed by the traditional Ad-hoc task. Only one participant keeps track of the offered tasks. This is likely due to the selected participants not being in the target group for CLEF, and also due to the CH evaluation lab only being launched this year and thus not having a long tradition and standardized framework like other domains. All participants are equally interested in evaluation of mature and new and experimental technologies. Since CHiC uses real Europeana data and queries, all participants were satisfied with the provided queries and data. No clear answer was given concerning the comparability of results since the majority had not participated in the lab and did not know about the results in detail.

2.3.6 Summary

The qualitative analysis of the use case showed strength and limitations which will be included in an updated use case. Due to the limited number of participants the results cannot be generalized but can be treated as recommendations for the further development of the use case framework.

2.4 Summary of the results - how do they affect the use case Framework?

The validity of the use case framework was controlled indirectly through combining the validation questionnaires from each use case domain and analyzing the typical answer patterns across the domains. The average scores and the variation in the scores between the use case domains were analyzed for the multiple choice questions concerning the use case description, system features, user features and session features.

The lowest average score for any question was 5.07 ("Have simplifications been made in the description of the system features?"), and the highest 8.46 ("Is the description of the use case readable?") on a scale of 1-10, 1 being the most negative and 10 the most positive score. Scores below 6.00 were interpreted to be more or less negative, and scores of 6.00 and above as more or less positive. The analysis focuses on the negative scores to identify possible problems and shortcomings of the use case framework. Due to the highly varied the responses, and the low number of respondents, the average scores can be misleading. Thus, to identify issues that were found problematic across the use case domains, even questions where at least one respondent from each domain gave a negative score were considered.

10 questions in the survey scored 6.00 or less, or had at least one negative score for each use case. Six of these questions were related to the description of the interaction pattern, three to the description of system features, and one question to the description of user features. The results are summarized in Table 1.

Question	Average score	Negative scores for each use case	Min./max. score Low/high median
1. Does the description of the use case consider variations of the flow?	6.00	yes	3/10 5/6,5 (M/IP)
2. Is it clearly stated where variations can occur?	5.23	yes	3/7 5/6 (CH/M)
3. Have simplifications been made in the description of the use case?	5.31	no	2/8 3/7,5 (CH/M)
4. Are the elements of interaction patterns well defined?	6.00	no	1/9 4/8 (CH/M)
5. Does the description of the session features cover all important features?	5.69	yes	3/9 5/6 (IP/CH, M)
6. Have simplifications been made in the description of the session features?	6.00	yes	2/9 4/8 (CH/M)
7. Are the boundaries of the system well defined?	6.00	yes	1/10 5,5/7,5 (IP/M)
8. Does the description of the system features cover all necessary aspects of information access systems?	5.53	yes	2/9 4,5/7 (IP/M)
9. Have simplifications been made in the description of the system features?	5.07	no	2/8 3/7,5 (CH/M)
10. Have simplifications been made in the description of the user features?	6.38	yes	2/10 6/7,5 (CH/IP, M)

Questions 1-6 in Table 1 are all (more or less) related to the description of the interaction sequence in the use case. The overly generic nature of the sequence of interactions was identified as the main problem for each use case; the interaction sequences were considered to simplify the real world interactions and to poorly present what variations could occur and where. Even if there was a lot of variation in the answers, and questions 3 and 4 were positively scored for the "visual clinical" use case, it seems obvious that the description of the interaction pattern was not properly supported in the use case framework, and needs to be revised.

Questions 7-9 in Table 1 relate to the description of the system features. Again the responses were very varied, but several implied the system features of the framework as something which should be looked over. The respondents maintained that not all the relevant aspects of the information access systems were described in the use cases, and that simplifications were made in their description. This might be related to the simplifications made in the description of the interaction sequences: if not all interactions are described, then not all system functionality will be included either. The framework however only explicitly describes the most basic system and interface functionality of information access systems. Adding more system features to the framework could be considered. When it comes to the system boundaries, the range of answers within the use cases was quite large, e.g. the scores for the "search for lecture material" use case ranged from 1-9 and for "visual clinical" use case from 4-10. While any interpretations will remain inconclusive, this might imply that it is difficult for stakeholders to understand what a system boundary is and what purpose it has. This should be made clearer in the framework.

Also, while the question was generally scored positively for all use cases, at least one respondent for each use case found that simplifications had been made in the description of user features (question 10 in Table 1). It's difficult to conclude whether this might imply problems in the framework, or simply reflect choices and simplifications made in the use cases. However, one responded mentioned that different dimensions of language skills could be better covered. Such issues are probably best handled at the level of individual use cases, as the framework cannot and does not aim to exhaustively cover the dimensions of all possible use case features.

3 Updated use case Framework and the checklists

3.1 Introduction

The results of the use case validation, and experiences from using the former version of the framework for the description of use cases and evaluation tasks, strongly suggested the necessity of a framework with enhanced support to guide the design of information access evaluation tasks. The revision focused on a number of aspects:

- Better over-all separation and description of the different factors that affects evaluation.
- Better description and guidelines for the design procedure
- A less labor-intensive procedure using checklists
- Better means to describe the user-system interaction
- Better means to model relevant features of the system

The revised version of the framework now assists evaluation design along a number of dimensions, held together in larger sets of features: *Background*, *Interaction*, *Interface and System*, and *Evaluation.*

The framework is called a "use case framework", as use cases, modeling system usage through the description of the user-system interaction, are at the very heart of the framework. It is in the interaction model that the constraints and demands related to the users and usage of systems meet the evaluation mechanisms: the characteristics of the

envisioned users, their tasks, contexts and environments all affect what interaction sequences are relevant to consider in evaluation. The background features cover these aspects. On the other hand (as this is an evaluation framework and not a system design methodology), the interface and system features of the operational systems evaluated, or of the experimental systems as defined in the experimental design, constraint the possible interaction patterns for a use case and thus limit the validity of the evaluation with respect to the users, search tasks, domains and environments covered. They are therefore described in the interaction and system model of the use case framework.

For each of these feature sets, a corresponding checklist has been formulated to support thinking about, designing, and documenting that aspect of information access usage or evaluation. The complete set of checklists, and the guidelines for using them, can be found in Appendices B-F. In the following, the different parts of the use case framework are discussed on a more general level.

3.2 Background Modeling

Individuals perceive their information needs in a subjective manner and the way they interact with information access systems depends on their goals, personal characteristics and attitudes. While some of the differences are genuinely individual, the users' group membership offers a strong signal of their possible needs and goals. *User role models* then define (abstract) user groups with respect to specific system usages. They are based on the tasks that users in specific roles are trying to accomplish while interacting with the system, but also describe the shared characteristics of those users, their interaction with the system and the information exchanged between the system and the users. The central *user role model* features include:

- User features, such as: user demographics (age, gender, education, social status); user knowledge and skills (with respect to the task, domain, system, language); physical characteristics ((dis)abilities); orientation and attitudes (towards the task, the system, co-searchers).
- Interaction features, related to the complexity, predictability and frequency of the interaction; locus of control of the interaction, and information flow direction.
- Information features, related to the volume and complexity of the information exchanged between the user and the system, as well as the clarity of the users' information needs.
- Users' primary success criteria, including: efficiency and effectiveness, system reliability and comprehensibility, actionability (does results enable taking intended action?).

Information access interactions are constrained by the activities that trigger them. A *domain model* captures the different constraints that govern a domain of activity: how the search behavior and goals of users are constrained by the activity at large (e.g. the "work" task) and the topic of interest; by the professional, private or social context of the activity (presence or absence of peers or collaborators while searching, sharing results with others); or by the characteristics of the data and repository accessed. A *domain model* may define e.g.:

- The cost of errors if search task is not duly completed (economic, social, societal, career, etc.).
- Time restrictions limiting the length of the interaction.
- Restrictions to accessing the contents of the repository (access rights, cost).
- Data and repository features, such media, genre, language quality and dynamics of the information/repository.

Different surroundings trigger different information needs and different interactions. The physical surroundings in which a user interacts with a system affect the search goals and the preferred way of interaction. An *operational environment model* depicts factors related

to the surroundings, mobility and locality of the users, distractions from the search interaction, and issues related to devices and network connections. The factors include, e.g.:

- Mobility and geo-position of the users
- Device and network restrictions (small screens, limited input ergonomics, high cost or low speed of data transfer)
- Distractions (interruptions, multiple parallel tasks, noise)

3.3 Modeling user-system interaction

The interaction between user and system is at the heart of our framework by modeling typical search sessions, and firmly connect user goals to interaction sequences. It is in the *interaction model* that the constraints and demands related to the users and usage of systems meet the evaluation mechanisms: on one hand, the users, domains and environments determine what interaction sequences and search goals are relevant to consider; and on the other, the evaluation mechanisms constrain the *interaction models* and thus limit the validity of an evaluation with respect to the users, search tasks, domains and environments covered. Therefore, the *interaction model* highlights the way the validity experiments gain from considering realistic interactions and search tasks is contrasted and adjusted to the requirements of feasibility in experimenting.

Correctly modeling the ways in which users interact with a system is essential for establishing the success criteria. An *interaction model* should cover the goals of the users, and depict the complexity of typical search sessions: search and result inspection strategies, result use, iterations of query reformulations, goal-orientation or randomness of the interaction. These aspects affect what results the users are likely to encounter and find relevant, given a certain time or effort of searching. They should therefore be reflected in both test collections and evaluation measures.

Use cases provide a useful framework for thinking about interaction in information access evaluation. There is no single established way of writing use cases, but use cases are typically organized around a *main success scenario* describing the simplest successful interaction sequence through the use case. The sequence is commonly presented as ordered steps, where each step describes one interaction between the user and the system. Main success scenario is complemented by a set of extensions that describe all the other possible interaction sequences through the use case, including any alternative user actions, exceptions and failures. A typical search use case may have a simple main success scenario (1. User types a query, 2. System shows results, 3. User clicks on a result 4. System presents result), but very many possible paths through the use case due to the high degree of freedom of user actions. Thus iterations of the different user actions in varying order need to be modeled through extensions.

The number of interaction sequences (main success scenarios and extensions) needed for describing most information access system usages is limited however: the number of identifiable user actions is not very high, and while the number of possible paths through the use cases might be overwhelming, the types of iterations of and switches between the actions are limited and thus possible to model through a limited number of interaction sequences and extensions.

The interaction sequences are here structured following [Wirfs-Brock 1993, Constantine 2006], by dividing the scenarios into *user intentions* and *system responsibilities* that show what the user aims to do in each step of the interaction and what system responsibilities relate to each user intention. Figure 1 depicts an example of a structured main success scenario for a use case for finding an illustrative image to insert in a blog post:

A goal in a use case refers to a concrete, immediate goal of a user interacting with the system, such as “inserting an illustration” in the above example. It defines the expected outcome of the interaction and thus introduces the immediate use of information as a factor affecting system evaluation criteria. A few goal categories with clear impacts on interaction

patterns have been recognized in previous studies, mainly based on analysis of web search logs [Broder 2002, Rose 2004]. They offer a solid starting point for considering goal categories, even if new categories to cover more varied usage and more specific goals may be needed. We separately define a second aspect of user goals, i.e., the type and amount of information looked for: single items or several items; ready answers, facts or notifications, or for topical content from which information can be extracted by the user.

insertingIllustration

USER INTENTION	SYSTEM RESPONSIBILITY
request illustration	
	show appropriate images
select image	
	show preview
confirm	
	insert image
	close

EXTENSIONS
browsingResults
reformulatingRequest

Figure 2. Example main success scenario.

3.4 Modeling interface and system

Interface design is an inseparable part of the interaction model, as even experiments where no users or interface designs are purposely included make assumptions concerning the user interface and system functionality: they are limited to specific request formulation functionality (e.g. typing keyword queries), and to specific result presentation functionality (e.g. ranked list of document titles) due to the way the experiment is set up. Such assumptions have a major effect on the applicability of the evaluation results and should not be overlooked.

From the use case example in Figure 1, three types of user actions and thus three groups of interface and system features may be identified: request formulation, result presentation (in two levels), and result use (inserting image). The interaction model then needs to be completed with a detailed (black-box) description of the interface features affecting the user's interaction with the system in these interaction points. The relevant aspects may include e.g.:

- Supported means for expressing requests: by querying or browsing; using different modalities; querying by examples or specifying queries by e.g. typing or humming.
- The granularity of the searchable information items: can queries target individual images, or (curated) collections or sets of images, or details in images, etc.
- Organization and presentation of the results: textual or visual results; thumbnails or full images, with context and copyright information, or without, etc.
- Result use such as manipulation, sharing, onsite consumption, exporting, ordering, etc.

3.5 Evaluation design

So, how do these models facilitate systematic construction of experiments based on rich models of users, domains, environments and interaction? The goal is a framework that can make explicit the functional requirements and success criteria of information access systems, and to connect them to benchmarking mechanisms, i.e., to the components of experimental settings and the criteria and metrics used for measuring system performance. Figure 2 depicts how the models are brought together:

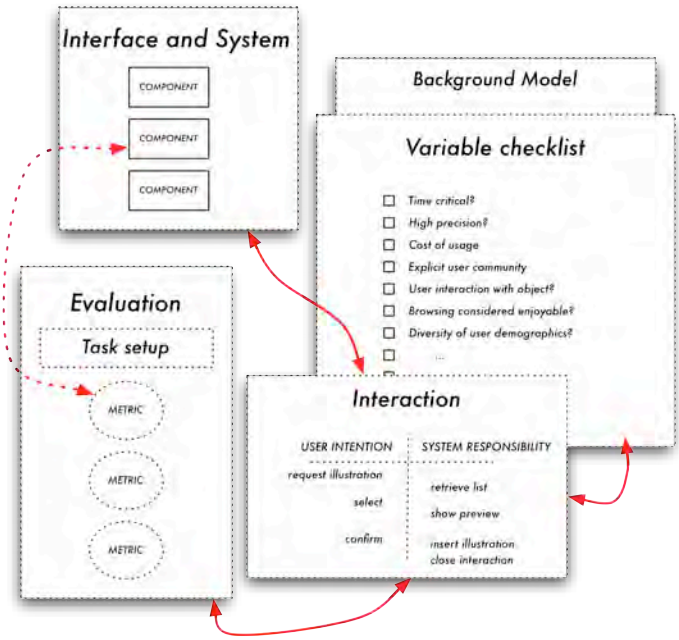


Figure 2. Bringing it all together.

The background models (user, domain, environment) collect the information needed for understanding the users' success criteria, and describe the preconditions of their interaction with the system: their abilities and preferences when it comes to formulating queries, inspecting results and interpreting and processing information. This information is then used in the design of experimental settings: for defining relevant information need (e.g. topics) and query types, the test data, relevance criteria and characteristics of the relevance assessors, interaction patterns that need to be modeled, and system interface features to cover.

The users' success criteria together with the interaction and interface models are needed for defining reasonable evaluation criteria based on the most important criteria for users, but also on what results they are likely to encounter when interacting with the system: Even if high recall is a prioritized success criterion for users, there is no point of basing evaluation on users ploughing through the entire result lists for one-shot queries if users typically search in sessions of several fast query reformulations and shallow result scans. The evaluation criteria as described through the interaction patterns can easily be operationalized in suitable metrics. Patience, time or cost parameters may be added into the standard metrics [e.g. Järvelin 2002, Moffat 2008], but probably yet new metrics need to be developed for measuring the quality of systems, given the varied success criteria of

users. The models and the process of mapping their features into experimental design can quite easily be formulated as easy-to-use checklists, similar to those used for documenting software system requirements, as implied in Figure 2.

A classic TREC-style batch experiment starts from topics which describe well formulated, clear, topical information needs. It extracts verbose keyword queries from textual topics descriptions. These are tested against static test collections with relevance assessments made by human expert assessors based on static relevance criteria. Evaluation is then over ranked lists of document pointers returned by the system and interaction with the system is modeled as sequences of one-shot queries and perusing the result list. The main success criterion used is effectiveness, as measured by MAP. This is potentially a useful experiment for evaluation of the quality of a ranking component in a search system for a use case describing some professional search tasks on the patent domain, where the cost of missing relevant documents may be high and users are thus willing to spend a lot of effort in formulating their queries and working down result lists.

It does not however capture the general success criteria for arbitrary other use cases. A system where users access information objects for entertainment with no clear task-related information need in mind and where the browsing itself is part of the use and enjoyment of the system and where one of the central goals of interaction may be participating in a community of users, exploring the organization of items into conceptual structures, understanding the intentions of the curator and other users of the resource, and possibly contributing to that community and the collection needs to be evaluated using entirely different metrics (e.g. [Murdock 2013]). Main success criteria for such system would be e.g. high levels of user engagement manifested as users returning to the site; long sessions with protracted browsing; user adoption of site terminology and categorization schemes; and numerous user actions, such as upvotes, comments, and share actions in response to returned item lists.

An evaluation of such a system might be based on a model of social interaction, with a test collection of linked data ranked by e.g. actionability - comments, votes, shares; "topics" describe unspecific and through the search session evolving information needs; requests reflect the users' evolving understanding of the vocabulary and conceptual model presented by the system. The system might be evaluated by average actionability of items it presents in response to user actions over explorative search sessions, by diversity and breadth of the presented set, and the burn-in rate in which the various dimensions of the system content are made available to the user.

The assumptions made concerning the users, their tasks, and interactions in experimental design have major effects on the realism and applicability of evaluation results. The use case framework supports re-establishing the links between the features of the real-world information access usages and the experimental settings pertaining to them, through broadening the focus from ranked system output to complete interaction sequences.

There are many different approaches to evaluation of information access systems. The suitable approach depends, in addition to the use case, on the target (component, complete service) and the perspective of the evaluation (goals of end users, goals of customers, and goals of service providers). Essentially, all types of evaluations benefit from carefully modeling the success criteria and interaction patterns for the evaluated systems. While focusing on improving the performance of isolated system components is motivated in some phases of technology development, such evaluations should not be agnostic about the end user benefits achievable (or not) by further improvements of the components.

3.6 Towards a Framework – use case relationships

Most experimental designs by necessity compromise between the breadth and the depth of their coverage: an experiment that aims to cover all users and all usages of a system, typically says very little concerning the systems' performance given any specific users or usages. On the other hand, the results from in-depth studies concerning the system usage

patterns of specific user groups working on specific tasks are most often difficult to generalize or to transfer to other situations.

The variation in the basic interaction sequences occurring in information access systems is however limited enough to be modeled through a set of predefined interaction sequence templates. Instances of information access usage can thus be described as use cases within a use case framework and related to other instances through their shared interaction sequences. A carefully constructed model of the relationships between the interaction sequences can then notably reduce the complexity of the "evaluation landscape" by bringing together the at first glance different information access use cases that ultimately are characterized by shared interaction patterns and goals and consequently, shared evaluation criteria.

Such a framework facilitates the generalization and re-use of evaluation results of the limited in-depth evaluations in other contexts and thus provides a platform on which evaluation criteria and evaluation results can be described, debated and validated. As more use cases are described, evaluated and validated within the use case framework, the knowledge of characteristics of use cases - with respect to evaluation and success criteria - will be enriched, and the connections between distinctive use case features and patterns of interaction and success criteria become clearer.

4 Final use cases and evaluation tasks

In a previous deliverable on this work package, deliverable 2.2 (Järvelin et al, 2012) we focused mainly on use cases. In this chapter, we focus mainly on evaluation tasks. Evaluation tasks typically benchmark or evaluate applications performing a service that is used by people in the context of a use case. For each evaluation task treated in this chapter, we briefly discuss its underlying use case first. Then the evaluation task is described. Finally, links between the evaluation task and the use case are discussed. This allows us to learn more about the validity of the evaluation tasks. Do these tasks relate to real-world tasks? Do benchmarking outcomes transfer to operational systems on real markets?

In a two-day project meeting in Gothenburg, Sweden, in December 2012, we filled in the new use case framework forms (discussed in Chapter 3) for each use case, and also the new evaluation task framework forms. In the ensuing discussion, incremental improvements for each of these forms were discussed. These improvements are already implemented in the version of the use case framework present in Chapter 3 in this deliverable. In the appendices G-K all filled-in forms from the Gothenburg meeting can be found. From time to time we will refer to these appendices in the discussions below. The idea of the meeting was to work out one or more evaluation tasks for one particular use case. As just discussed, the goal of this exercise is to learn more about the validity of the evaluation tasks. The forms can be used as an aid in this process. The use case forms contain space to refer to properties of the evaluation activity and vice versa. The evaluation task summary form is a one-stop place where an evaluation task may be decomposed in its components, each of the components can be related to features of the use case, and any simplifications in the way the component was implemented may be indicated.

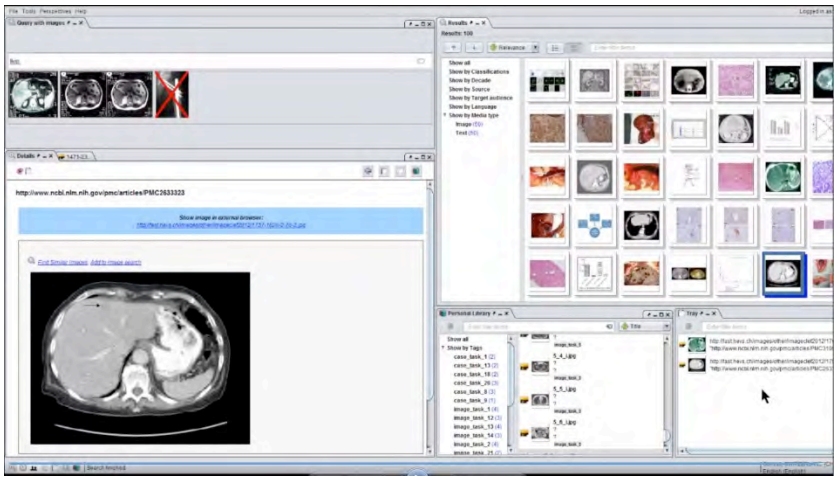
In the sections below we will see how researchers made use of different forms – some researchers preferring one, some preferring to use another – to arrive at a high quality evaluation task description. For each use case domain, we cover one use case and one or more evaluation tasks. Two use case domains deserve a separate introduction. First, the online reputation management use case and evaluation tasks are discussed only briefly, as the main effort here is part of the Limosine EU project (http://limosine-project.eu). However, a section on validation of the evaluation tasks is added here. The reputation use case was written based on the updated use case framework discussed in Chapter 3, and was validated in a stakeholder interview following the updated use case framework structure. This validation was different from the validation efforts reported in Chapter 2 and is therefore

discussed here. Second, the enterprise search domain covers an evaluation activity using the black box evaluation framework detailed in deliverable 4.2 [Garcia Seco de Herera et al, 2012]. It represents the biggest divergence from Cranfield style evaluation in the evaluation activities we cover in this deliverable.

4.1 The medical domain

4.1.1 The visual clinical decision support for medical diagnosis use case

This use case deals with visual information connected with text in the radiology domain in order to help clinicians find medical cases/images similar to the one under observation. This supports a clinician's decision making during medical diagnosis using medical images and text describing the case under observation as queries in biomedical literature.



4.1.2 The American medical informatics association medical task

In 2013, the ImageCLEF medical task will, for the first time, organized a workshop outside of Europe; at the annual American Medical Informatics Association (AMIA) meeting (http://www.amia.org/amia2013) in the form of a workshop. This medical task consists of four subtasks in 2013. The use case framework and evaluation task framework forms in Appendix C were filled in with two of these subtasks in mind: the ad-hoc image based retrieval task and the more complicated case based retrieval task. Both tasks have the same underlying use case, but different units of retrieval. We base our description of both tasks on the task description for the 2013 task (http://www.imageclef.org/2013/medical) and on the overview paper of the 2012 task [Müller et al, 2012], in order to give a complete description of the current state of the tasks.

4.1.2.1 Datasets

ImageCLEFmed 2013 uses the same database that was used in 2012. In 2012, a larger database than 2011 was provided using the same types of images and the same journals. The database contains over 300,000 images of 75,000 articles of the biomedical open access literature that allow free redistribution of the data. The ImageCLEF database is a subset of the PubMed Central database containing in total over 1.7 million images. PubMedCentral contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals.

4.1.2.2 Ad-hoc image-based retrieval topics:

This is the classic medical retrieval task, similar to those in organized in 2005-2012. In the AMIA 2013 medical task participants will be given a set of textual queries with 2-3 sample images for each query. Queries are in English, and translations in Spanish, French or German are also provided. The queries will be classified into textual, mixed and semantic, based on the methods that are expected to yield the best results. The unit of retrieval is an image. The topics for the image–based retrieval task are based on a selection of queries from search logs of the Goldminer radiology image search system [Tsikrika et al, 2012]. Only queries occurring more in the logs than a certain threshold are considered as candidate topics for this task, resulting in about 200 topics. Then, a radiologist assesses the importance of the candidate topics, resulting in about 50 candidate topics which are then further reduced if there are not at least a few relevant items in the database. The resulting queries are then distributed among the participants and example query images are selected from a past collection of ImageCLEF.

4.1.2.3 Case based retrieval topics

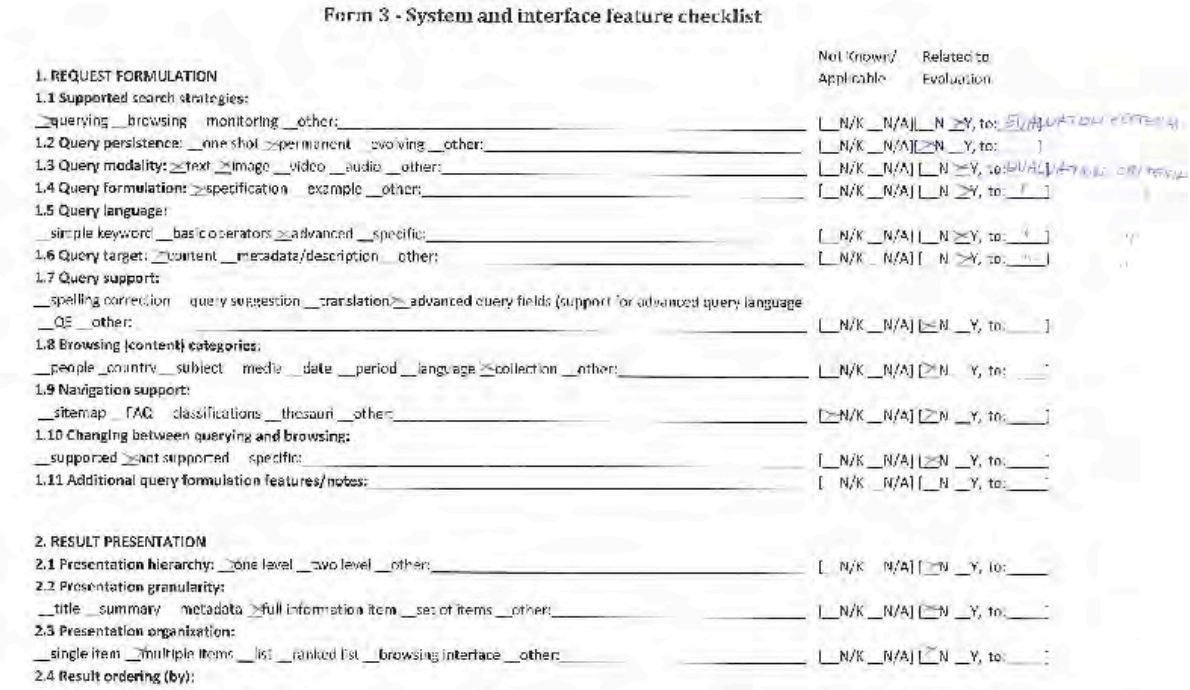
The case–based retrieval task was first introduced in 2009. This is a more complex task but one that we believe is closer to the clinical workflow. In this task, case descriptions with patient demographics, limited symptoms and test results including imaging studies are provided (but not the final diagnosis). The goal is to retrieve cases including images that a physician would judge as relevant for differential diagnosis. Unlike the ad–hoc task, the unit of retrieval here is a case, not an image. The topics are created from an existing medical case database. Topics include a narrative text and several images.

4.1.2.4 Relevance judgments

The relevance judgements will be performed with the same on–line system as in 2008–2012 for the image–based topics as well as case–based topics. For the case-based topics, the system displays the article title and several images appearing in the text (currently the first six, but this can be configured). Judges were provided with a protocol for the process with specific details on what should be regarded as relevant versus non–relevant. A ternary judgment scheme will be used, wherein each image in each pool is judged to be “relevant”, “partly relevant”, or “non–relevant”. Images clearly corresponding to all criteria are judged as “relevant”, images for which relevance could not be accurately confirmed are marked as “partly relevant” and images for which one or more criteria of the topic are not met are marked as “non–relevant”. Judges are instructed in these criteria and results are manually verified during the judgment process. As in previous years, judges are recruited by sending out an email to current and former students at OHSU’s (Oregon Health and Science University) Department of Medical Informatics and Clinical Epidemiology. Judges, primarily clinicians, are paid a small stipend for their services. It is the goal to have many topics judged by two or more judges to explore inter–annotator agreement and its effects on the robustness of system ranking.

4.1.2.5 Evaluation metrics

Evaluation metrics for both the image-based and the case-based retrieval task are standard information retrieval metrics. The assumption is that the user is helped with a ranked list of retrieved items, where the items which are deemed most relevant are listed at the top. The primary metric is mean average precision (MAP). This is a well-studied and well-understood metric that rewards high precision (low number of non-relevant results) as well as high recall (a large fraction of relevant results is located). To understand better what is going on, different metrics are also published: GM_MAP, bpref, P10 and P30. For understanding of these metrics, we refer the reader to [Sanderson, 2010].



Form 3 - System and interface feature checklist

	Not Known/ Applicable	Related to Evaluation
1. REQUEST FORMULATION		
1.1 Supported search strategies:		
_querying __browsing __monitoring __other:	[__N/K __N/A]	[__N >Y, to:]
1.2 Query persistence: __one shot >permanent __evolving __other:	[__N/K __N/A]	[>N __Y, to:]
1.3 Query modality: >text >image __video __audio __other:	[__N/K __N/A]	[__N >Y, to:]
1.4 Query formulation: >specification __example __other:	[__N/K __N/A]	[__N >Y, to:]
1.5 Query language:		
__simple keyword __basic operators >advanced __specific:	[__N/K __N/A]	[__N >Y, to:]
1.6 Query target: >content __metadata/description __other:	[__N/K __N/A]	[__N >Y, to:]
1.7 Query support:		
__spelling correction __query suggestion __translation >advanced query fields (support for advanced query language __QL __other:	[__N/K __N/A]	[>N __Y, to:]
1.8 Browsing (content) categories:		
__people __country __subject __media __date __period __language >collection __other:	[__N/K __N/A]	[>N __Y, to:]
1.9 Navigation support:		
__sitemap __FAQ __classifications __thesauri __other:	[>N/K __N/A]	[>N __Y, to:]
1.10 Changing between querying and browsing:		
__supported >not supported __specific:	[__N/K __N/A]	[>N __Y, to:]
1.11 Additional query formulation features/notes:	[__N/K __N/A]	[__N __Y, to:]
2. RESULT PRESENTATION		
2.1 Presentation hierarchy: __one level __two level __other:	[__N/K __N/A]	[>N __Y, to:]
2.2 Presentation granularity:		
__title __summary __metadata >full information item __set of items __other:	[__N/K __N/A]	[>N __Y, to:]
2.3 Presentation organization:		
__single item __multiple items __list __ranked list __browsing interface __other:	[__N/K __N/A]	[>N __Y, to:]
2.4 Result ordering (by):		

Figure 4. First page of Form 3 – System and interface features for the Medical use case.

4.1.3 Discussion on relation between use case and evaluation task.

Strong points of the medical use case and evaluation tasks combo are the involvement of medical experts in the creation of both topic and relevance judgments. A retrieval test collection typically consists of queries, documents and relevance judgments. In the image-based and case-based medical task test collections queries are clearly representative of real life queries since they are first mined from the log files of an operation search engine and subsequently scrutinized by medical experts. The document collection is actually a real and high quality collection. Finally, relevance judgments are representative of real-life relevance since they are performed by medical experts. The only limitation is the current standard of performance of retrieval algorithms. Because pooling is used in the relevance judgment process, images or cases which are not retrieved by any retrieval algorithm will not be judged. This is a well-known but as of yet unsolved problem in the information retrieval community. As document collections grow, the problem becomes more apparent, and the document collection used for the image-based and case-based medical tasks is growing every year. A large variety of retrieval algorithms will ensure a diverse pool at least, mitigating the problem. A test collection is designed with a task in mind. Of the two subtasks we treat in this deliverable, the case-based retrieval task is believed to be closest to everyday practice of clinicians. Still, systems performing the image-based retrieval task have merit, e.g. as a component of a system performing the case-based retrieval task.

4.2 The intellectual property domain

4.2.1 The claim to validity use case

A typical scenario for the Claim to Validity Use Case is the following: A professional searcher receives a patent application document and is told to find other documents that may invalidate the claims in the application. The searcher studies the application and, from a patent repository, pulls out related patent documents (applications at other patent offices, previous patents of the applicant, etc.). Using the text of the claims and few other terms extracted from the documents, he formulates queries which he submits to a search system specialized on patent data. He examines the documents returned in the result list, marks some with various degrees of relevance, marking also the passages relevant to the set of

claims in the given application document. When the searcher has gathered enough data to support taking a decision, he will draft a report stating the results of his search. The report contains the list of documents that he considers relevant to the validity of the given set of claims. The documents in the report are listed together with their degree relevance and with concrete references to images, tables, pages and line numbers where the text in the relevant document is the most pertinent to the set of claim in the given patent application document. In one sentence, the Claim to Validity Use Case may be summarized as: Given a set of patent claims we want to find passages of text relevant to this set which may invalidate what is claimed.

4.2.2 The claims to passages evaluation task

In one sentence, the Claims to Passages evaluation task can be summarized as:

Given a claim, retrieve relevant documents in the collection and mark out the relevant passages in these documents. As topic background data, which can be used in query generation, the patent application document in which the claims occur together with any family members of this application document are made available.

The 'Claims to Passages' evaluation task is closely related to the 'Claim Validity' Use Case.

The topics of the evaluation task are claims occurring in patent application documents. The patent documents are actual patent applications already processed at and published by the European Patent Office (EPO). This means that each patent application comes with a search report created by an expert patent examiner at the EPO. The search report contains the list of documents which the examiner considered as relevant to the given patent application. Each document entry in the list carries a relevancy degree, indications to which particular claim in the application document the relevant document pertains to, and often enough indications to which page(s), images, tables, etc. are particularly relevant to the particular claims. Since the patent application documents are in one of the three EPO official languages (English, German, French) the topics of the evaluation task may also be in any of these three languages. The choice of patent documents used for topic extraction is done firstly by the number of entries in the search reports (also called `citations') and by the type of these citations - highly relevant, marginally relevant, relevant in combination with, etc.

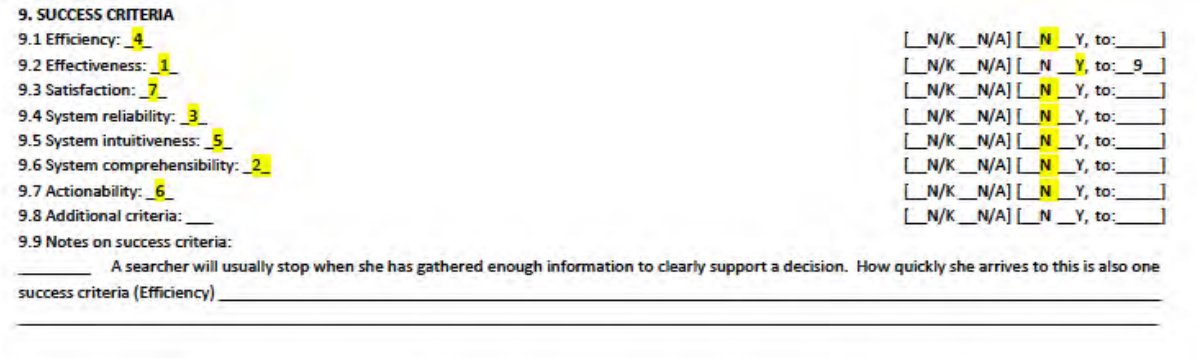
The data corpus that is used by this evaluation task includes the patent documents published by the European Patent Office until 2002. The patent collection (known as the CLEF-IP collection) stores the patent documents as XML files. There are over 3 million XML files in this collection. The patent documents that are used as sources for the task's topics are in the same XML format and _not_ part of the CLEF-IP corpus.

The concrete representation of both the topics and the relevance assessments for these topics involve the XML representation of the patent documents and XPaths that identify the parts of the document that is to be used. The task's evaluation topics are XPaths that contain claim text. Queries to retrieval systems may process the text as seem fit for best representing the information need. Eventual previous applications of the topic patent to other patent offices are also made available and can be used for term extraction. The relevance assessments for a topic contain a list of patent document identifiers (which is also the XML file name without the extension) and XPaths in these documents that identify relevant text. In one relevant document more XPaths can be relevant.

The relevance assessments for the `Claims to Passages' evaluation task are based on the same search reports that were used to choose them. Since the search reports are available only in a PDF format we have to manually extract the information of interest out of them and identify the relevant text in the documents within the CLEF-IP collection, then save the respective XPaths in the relevance assessment files (qrels).

The results given by retrieval systems are compared against the documents and XPaths in the qrels. Based on this comparison we compute two main measures at both document and passage levels: We look first if the documents retrieved match those in the qrels (document level) computing Precision and Recall; Then we look, within the retrieved documents, at the

proportion of relevant text retrieved (as XPaths) computing Precision and Recall at passage level.



9. SUCCESS CRITERIA

9.1 Efficiency: _4_	[__N/K __N/A] [__N __Y, to:_____]
9.2 Effectiveness: _1_	[__N/K __N/A] [__N __Y, to:__9__]
9.3 Satisfaction: _7_	[__N/K __N/A] [__N __Y, to:_____]
9.4 System reliability: _3_	[__N/K __N/A] [__N __Y, to:_____]
9.5 System intuitiveness: _5_	[__N/K __N/A] [__N __Y, to:_____]
9.6 System comprehensibility: _2_	[__N/K __N/A] [__N __Y, to:_____]
9.7 Actionability: _6_	[__N/K __N/A] [__N __Y, to:_____]
9.8 Additional criteria: ___	[__N/K __N/A] [__N __Y, to:_____]

9.9 Notes on success criteria:

_____ A searcher will usually stop when she has gathered enough information to clearly support a decision. How quickly she arrives to this is also one success criteria (Efficiency) _____

_____

Figure 5. Success criteria for the IP use case: Form 1 – Background features.

4.2.3 Discussion on the relation between use case and evaluation task.

We discuss some links that were marked on the use case and evaluation task forms in Appendix H. Then we discuss some lines of improvement we plan to pursue in future work.

1. F4. 2.1 -> F1. 2.1, 2.4

The topics are representative of the real search situations where the information need:

- is clearly expressed (`find text that invalidates a claim')
- is multilingual (the relevant text may be in another language than the language of the claim text)
- refers to specific domains of knowledge, domains in which the searcher is highly educated

2. F4. 2.1 -> F2. 2.1; F4.3 -> F1. 5.6; F4.10 <-> F1. 9

The results shown to a searcher in a real-life situation is a list of documents where relevant passages are (ideally) marked out. The searcher will browse through this list, marking relevant content and/or documents, will refine the search query, and continue until satisfied with the list of selected results. The results in our ET is also a list of documents with the relevant passages marked out. We consider this list to be the final list of documents selected in a search session.

3. F4.10 -> F3.3

The time of IP experts is very dear, so one input to the level of satisfaction of an expert IP searcher is to find enough relevant results (recall) and find them early (precision). In the ET we measure recall and precision both at document and passage level.

For the use case discussed above, the final list of relevant documents and mark-ups are stored in a search report, which the ET uses for its relevance assessments. However, any thorough search activity in the IP domain happens in cycles and one line of future work is to include this cyclic attribute in the ET. There are, though, certain obstacles, like obtaining intermediate lists of results, which may not be easily overcome.

Another line of future work is to compare how the one-value metric results of automatic runs match to the expert's real expectations from a retrieval system specialized on patent data.

4.3 The cultural heritage domain

4.3.1 The search for cultural heritage material use case

A history professor, searches various images of soldiers in high-definition and without license restrictions in a digital library in order to prepare a presentation for his lecture. He uses Europeana, a large-scale reference database 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 main success scenario could be described as the following basic flow of interaction. The history professor selects Europeana as a 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 possibly 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 in order to view and save the original object. In the end he creates a collection of images. 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 behavior.

The goal of this task is an overview / list of images related to the topic for further use. 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.

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. For the structural description of the use case framework see Appendix I: Use Case and Evaluation Task Forms for Cultural Heritage.

Form 2 – Interaction and goals – Search for cultural heritage

1. USE CASE NAME AND SUPPORTED USER ROLES

1.1 Name: Search for cultural heritage material

1.2 Supports (user roles): Searcher, browser, flaneur

2. USER GOALS	Not Know/ Applicable	Related to Evaluation
2.1 Type of information: __single fact/answer/notification __collection of facts/answers/notifications __single item (e.g., document) x collection of items __other:	[__N/K __N/A]	[__N x Y, to:___8]
2.2 Type of goal: x viewing x exporting __navigating __ordering/buying (transactional) __manipulating __surfing __other:	[__N/K __N/A]	[__N x Y, to:___8]

3. USE CASE RELATIONSHIPS

3.1 Specializes:	[__N/K x N/A]	[__N __Y, to:____]
3.2 Extends:	[__N/K x N/A]	[__N __Y, to:____]
3.3 Uses: simple ad-hoc search, simple browse	[__N/K __N/A]	[__N x Y, to:___7]
3.4 Resembles: simple ad-hoc search, simple browse	[__N/K __N/A]	[__N x Y, to:___7]

4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

Search for CH Material	EXTENDS: Ad-hoc search
USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction	
Type query	Present result list
Choose facet	Reduce result list according to facet characteristic
Browse result pages	Enable paging of result list
View individual object	Open individual object page
Click related object	Open individual object page
Click external link	Enable link tracing
	Close (Use Case Ends)
EXTENSIONS	

Figure 6. Form 2 – Interaction and goals for the Cultural heritage use case.

4.3.2 The multilingual ad-hoc search task

The task “multilingual ad-hoc search” is organized within the CLEF (Cross-Language Evaluation Forum) Cultural Heritage in CLEF (CHiC) track.

This task is a standard ad-hoc retrieval task, which measures 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, the multilingual ad-hoc retrieval task requires participants to submit as many relevant documents from the whole multilingual Europeana collection as possible, meaning, the documents can be in any language the collection provides.

Data: CHiC uses a static (canned) version of the whole Europeana index (23 million multilingual metadata objects describing texts, images, audio and video files). The data appears exactly as in the portal (same information content), however, during the duration of the evaluation task, the collection is not updated or changed. For processing purposes, the Europeana collection has been divided into sub-collections according to metadata field language – 13 sub-collections have been formed.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. The 50 short topics in title-format only (e.g. "Eiffel tower") reflect real expressed user needs as represented in Europeana (taken from actual query logs). The topics for CHiC multilingual ad-hoc will be in English.

Expected results: Participants are expected to submit relevance-ranked result lists for all 50 topics in a ranked list format using documents from the multilingual Europeana collection.

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).

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. In Table 4.1 below we present another schematic overview of the evaluation task. We break the task down by its components, discuss the relations of each component to the use case and discuss how we chose to implement the component.

Table 4.1: Filled in evaluation task summary form for the multilingual ad-hoc search task and the search for lecture material use case.

Component	Use case features considered	Instantiation of the component
1. Test subjects	./.	No actual users will participate.
2. Topics	Real-life topics will be used	Information needs from actual Europeana users (gathered from log files and elaborated through group discussions) will be used
3. Requests	Real-life requests will be used	Requests from actual Europeana users (gathered from log files) will be used
4. Data	The information system and data as described in the use case will be used.	The data is only slightly changed by creating language-dependent subcollection. The actual metadata format as in the real collection will be used.
5. Ground truth creation	Assessors familiar with cultural heritage content will judge the relevance of objects as if they are real Europeana users.	Close approximation of actual users will be used for intellectual, manual relevance assessments.
6. Result presentation	Europeana uses a 3x4 matrix of thumbnails + object titles to present in a ranked list. The evaluation experiment shares the ranked list but the matrix is not used for relevance assessments.	Participants must present a ranked list of Europeana objects for any request.
7. Interaction	No actual user interaction. A single search interaction is simulated by using individual requests and ranked lists.	No actual user interaction. Requests are automatically submitted to an information system, ranked lists are assessed by assessors.
8. Result use	Not considered	Not considered
9. Evaluation criteria	Out of the two relevant evaluation criteria (effectiveness and satisfaction), only effectiveness is considered.	The goal of the evaluation task is to measure system effectiveness in finding relevant objects for a given request.

10. Metrics	Appropriate IR effectiveness measures are used.	MAP, P@k

4.3.3 Discussion on the relation between use case and evaluation task

The evaluation task CHiC multilingual ad-hoc search was developed based on one interaction component of the "Search for cultural heritage material" – the initial search and result list viewing interaction. While further interactions like filtering or related object browsing are not considered, the initial search (user types in a query and reviews the result list) can be considered as the primary interaction in the use case. Since the evaluation task uses both data, topics and requests from the information system described in the use case, it is considered a real-life application of the use case. The use case is not as well represented in the ground truth creation (i.e. relevance assessments), because real Europeana users are only involved insofar the human assessors also – coincidentally – use the Europeana portal. The human assessors are, however, familiar with cultural heritage content and the functionalities of digital libraries so that the assessments will be appropriate for the more abstract case. More explicit links between use case and evaluation task can be found in Appendix I.

4.4 The online reputation management domain

We will keep the description of this use case and evaluation task brief, as the work is mainly carried out in the Limosine project (http://www.limosine-project.eu/about).

4.4.1 The online reputation management use case

In the context of ORM, monitoring refers to a constant (e.g. daily) scrutiny of online (and, in particular, social) media searching for information related to the entity. It focuses on the opinions and news related to a given company and aims at early detection of any potential menace to its reputation, that is, issues and opinions that could damage the company's public image. That implies a frequent inspection of the most recent online information. Microblogs and, especially, Twitter, are key sources for this task (Amigó et al, 2012)

4.4.2 The online reputation management monitoring task

In the monitoring task, systems receive a stream of tweets containing the name of an entity, and their goal is to (i) cluster the most recent tweets thematically, and (ii) assign relative priorities to the clusters. A cluster with high priority represents a topic which may affect the reputation of the entity and deserves immediate attention (Amigó et al, 2012).

4.4.3 Discussion on the relation between use case and evaluation task

A strong point of the RepLab campaign is that professional reputation managers from Llorente & Cuenca were involved in the design of the tasks, and also created the ground truth for the task. Therefore, the assessors were perfectly representative for the end user population. In our discussion on the validation of the online reputation management use case in the corresponding subsection of chapter 2, we discussed other links between the use case and the monitoring evaluation task.

4.4.4 Reputation management – testing the new framework

In the Limosine project (http://www.limosine-project.eu/) the online reputation management use case plays a central role. At CLEF 2012 they organized the successful RepLab benchmarking campaign (Amigó et al, 2012). This campaign featured two evaluation tasks: the monitoring task and the profiling task. The profiling task was a simplified task, accessible to participants, and many groups joined it. The monitoring task was a task closer to the everyday practice of reputation managers. In fact, reputation managers of Llorenta & Cuenca participated in the design of both tasks. However, the monitoring task was much more complex and fewer groups joined. In the next edition of RepLab the monitoring task

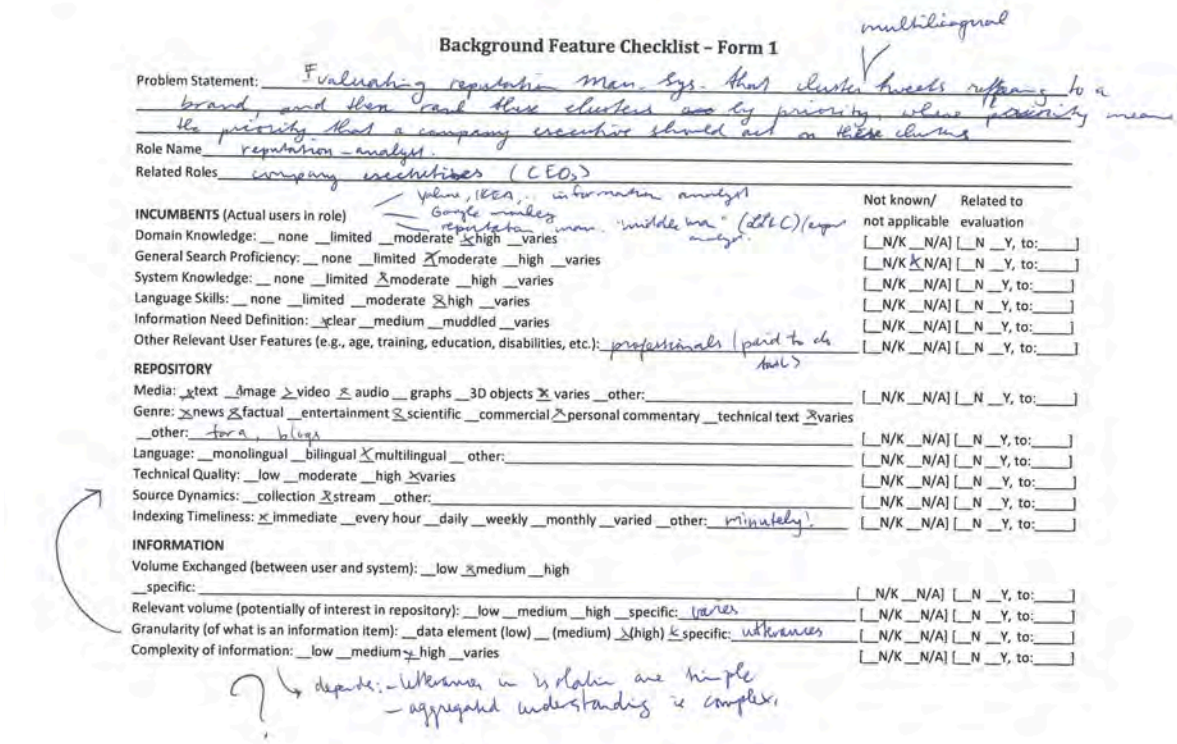
will be the only task. Baseline systems for all necessary components of a system performing the monitoring task will be provided by the organization, making participation easier.

In the Promise project, we interviewed Magnus Sahlgren from Gavagai (http://www.gavagai.se/), a Stockholm based company also involved in online reputation management. Note that we went a step further than only validating if a use case description corresponds to real life practice. We went there also with an evaluation task in mind and discussed whether or not the evaluation task corresponds to the use case. The timing of this interview was after the development of the current version of the use case framework, and therefore the validation protocol which was designed for the deliverable 2.2 (Järvelin et al, 2012) did not apply. Instead, we filled in the use case framework together, starting with the profiling task description. We used the use case framework as a starting point to discuss the relation between the RepLab benchmarking tasks and day to day practice of reputation management practitioners. We quickly discovered that indeed the profiling task has no obvious relationship to any task in the wild. Therefore we switched to the monitoring task. As we progressed in our discussion, Magnus realized that this task and the underlying use case are in fact spot on. Customers in his experience are not satisfied with nice graphs which summarize data they are interested in. Instead, actionable intelligence is the key. In the monitoring task, systems provide a clustering of tweets by topic, and rank these clusters by priority. A high priority entails that action needs to be taken on this cluster: tweets in this cluster should be examined for the harm they may do brand reputation.

We did find some points on which the realism of the online reputation management monitoring task might be improved. Magnus noted that in real life it is not enough to monitor Twitter streams. Blogs and Fora are examples of other channels which are essential to monitor. Also, it is interesting to note that observing can change reality here: if people do not want a conversation to be monitored, they will flock from monitored media. Another crucial aspect of the online reputation management monitoring use case is the real time constraints. At Gavagai indexes are refreshed continuously, minute by minute. Tweets are very short by nature. In general, however, it is not enough to analyze text on document level. Instead, at Gavagai, the unit of interest is an utterance.

Finally, for the monitoring task no suitable evaluation metrics existed. The organizers of the task therefore devised a “new and exciting” metric for the task. One line of future work could be to scrutinize to what extent this metric corresponds to real life requirements. An interesting example question is: are all mistakes equally harmful? For example, can the metric be adjusted to heavily penalize the failure of ringing an alarm for certain tweet clusters?

For the filled in use case and evaluation task forms see Appendix J: Use case and evaluation task forms online reputation management.



Background Feature Checklist – Form 1

Problem Statement: Evaluating reputation man. sys. that cluster tweets referring to a brand and then rank these clusters ~~are~~ by priority, where priority means the priority that a company executive should put on these clusters

Role Name: reputation – analyst.

Related Roles: company executives (CEOs)

	Not known/ not applicable	Related to evaluation
INCUMBENTS (Actual users in role)		
Domain Knowledge: __ none __limited __moderate X high __varies	[__N/K __N/A]	[__N __Y, to:____]
General Search Proficiency: __ none __limited X moderate __high __varies	[__N/K X N/A]	[__N __Y, to:____]
System Knowledge: __ none __limited X moderate __high __varies	[__N/K __N/A]	[__N __Y, to:____]
Language Skills: __ none __limited __moderate X high __varies	[__N/K __N/A]	[__N __Y, to:____]
Information Need Definition: X clear __medium __muddled __varies	[__N/K __N/A]	[__N __Y, to:____]
Other Relevant User Features (e.g., age, training, education, disabilities, etc.): professionals (paid to do task)	[__N/K __N/A]	[__N __Y, to:____]
REPOSITORY		
Media: X text __image X video X audio __graphs __3D objects X varies __other:____	[__N/K __N/A]	[__N __Y, to:____]
Genre: X news X factual __entertainment X scientific __commercial X personal commentary __technical text X varies __other: forums, blogs	[__N/K __N/A]	[__N __Y, to:____]
Language: __monolingual __bilingual X multilingual __ other:____	[__N/K __N/A]	[__N __Y, to:____]
Technical Quality: __low __moderate __high X varies	[__N/K __N/A]	[__N __Y, to:____]
Source Dynamics: __collection X stream __other:____	[__N/K __N/A]	[__N __Y, to:____]
Indexing Timeliness: X immediate __every hour __daily __weekly __monthly __varied __other: minutely	[__N/K __N/A]	[__N __Y, to:____]
INFORMATION		
Volume Exchanged (between user and system): __low X medium __high __specific:____	[__N/K __N/A]	[__N __Y, to:____]
Relevant volume (potentially of interest in repository): __low __medium __high __specific: varies	[__N/K __N/A]	[__N __Y, to:____]
Granularity (of what is an information item): __data element (low) __ (medium) X (high) X specific: references	[__N/K __N/A]	[__N __Y, to:____]
Complexity of information: __low __medium X high __varies	[__N/K __N/A]	[__N __Y, to:____]

depends: – references in isolation are simple – aggregated understanding is complex.

Figure 7. First page of the Form 1 (Background feature checklist) for the Reputation management use case.

4.5 The enterprise search domain

4.5.1 The enterprise search use case

In enterprise search, the end user has landed on the site of some company or institution. Typically, he or she is looking for information about products or services of this organization. There will be several ways to do that: via the navigation links offered by the site, or via some search functionality. The use case covers both of these scenarios, even though the focus remains mostly on search functionality.

4.5.2 The black box evaluation task for enterprise search

The evaluation is performed using a black box application evaluation methodology performed as a guerrilla evaluation campaign. This black box methodology was covered primarily in deliverable 4.2 [Garcia Seco de Herera et al, 2012]. Here, applications are treated as black boxes and cannot be fully instrumented to test single system components in isolation. Instead, the application as a whole is under scrutiny, considering the interplay between core IR systems, user interfaces, underlying data and configurations. “Guerilla evaluation” means that live applications are evaluated without the knowledge of companies whose applications are evaluated. One could say that an end user in the black box evaluation framework is modeled through the combination of the test protocol and the tester. This setting is far away from the usual Cranfield style benchmarking campaigns and as such it constitutes a test of the use case and evaluation task framework: will they be useful even for very different use case and evaluation scenarios? In the Table below we summarize a typical black box evaluation activity.

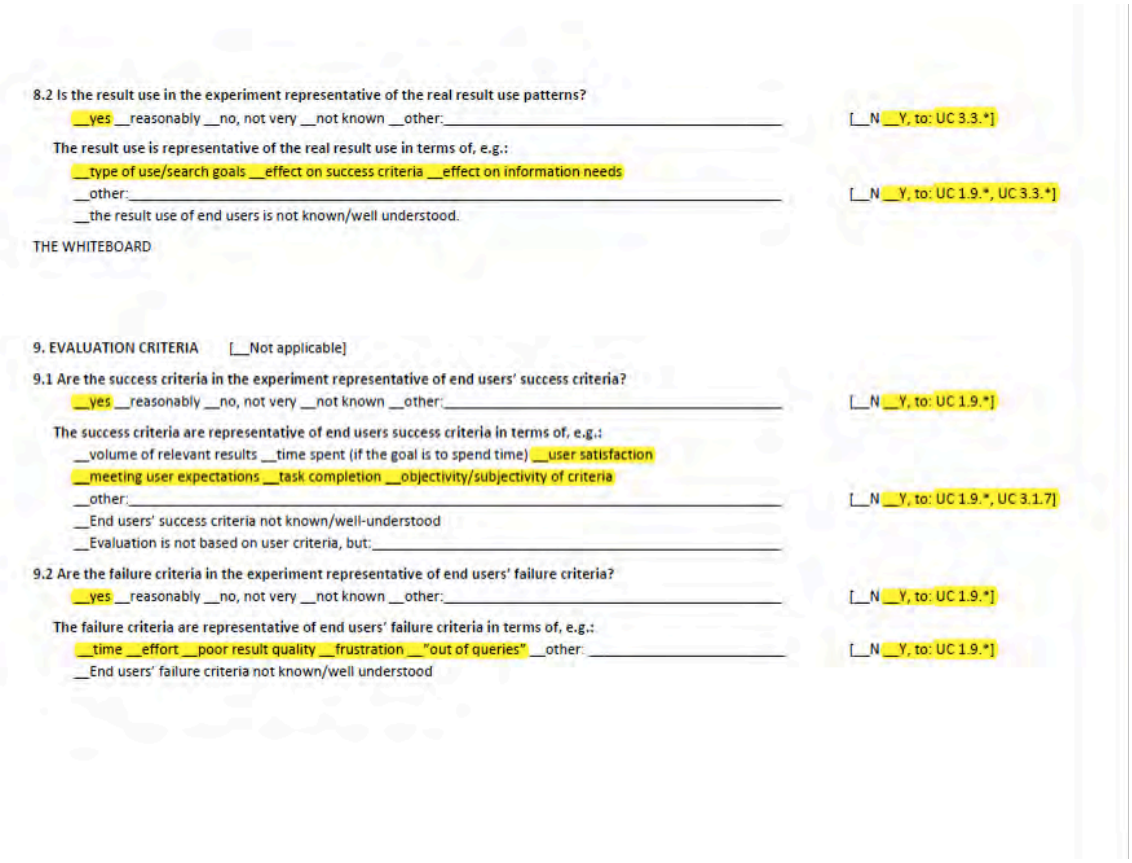
Table 4.2: Filled in evaluation task summary form for the black box evaluation task and the enterprise search use case.

Component	Use case features considered	Instantiation of the component
1. Test subjects	All incumbent features (1.2.*)	Test scripts which model prototypical users' behavior. Testers only execute the scripts and are not the modeled subjects.
2. Topics	All user goal features (2.2.*)	Information needs are described as abstract templates in test scripts and are made concrete by testers based on the tested application's domain.
3. Requests	Incumbent domain knowledge (1.2.2), Incumbent general search/system proficiency (1.2.3), Access restrictions (1.7.3), Request formulation features 3.1.1-6,8: supported search strategies, query persistence, query modality, query formulation, query language, query target, browsing categories	Queries are formulated by testers on the fly based on the test scripts' descriptions.
4. Data	All repository features (1.3.*)	Live application data, which is also subject to test.
5. Ground truth creation	n/a	Ground truth differs from each application and test and must be assessed as best as possible by testers and is therefore not comparable to the ground truth in typical experiments.
6. Result presentation	Repository media (1.3.1), Repository granularity (1.3.2), all result presentation features (3.2.*)	Live applications' result presentation is used and tested.
7. Interaction	Origin of user input (1.4.1), clarity of information need (1.4.2), flow direction (1.4.3), all interaction features (1.5.*), Navigation support (3.1.9), changing between querying/ browsing (3.1.10)	Test scripts assume natural interaction with live applications and several aspects thereof are tested.

8. Result use	All result use features (3.3.*), all success criteria (1.9.*)	The results generated by tested applications are assessed in terms of satisfying success criteria of tests.
9. Evaluation criteria	All success criteria (1.9.*), query support (3.1.7)	Test criteria and scripts are modeled according to expected user needs and behavior, therefore the overall success criterion is meeting user expectations.
10. Metrics	Cost of errors (1.7.1), all success criteria (1.9.*)	User perception of an application.

4.5.3 Discussion on the relation between use case and evaluation task

A strong point here is that real life, operational systems are evaluated. Another strong point is that such applications are evaluated in a very comprehensive way, taking into account the user interface, index freshness, quality of metadata, quality of ranking, and so on. Users are modeled in an interesting way, via an elaborated test protocol and a restricted tester, acting almost like a robot. While this may be a simplification of reality, it enhances repeatability of evaluation experiments. The black box evaluation framework may be adapted to different use case domains by manipulating the weights for individual tests, or even categories of test. Also, different queries or tests may be devised for different use case domains. The above presentation was discussed in the context of the enterprise search domain. Exploring the evaluation of systems in other domains is an interesting line of future work.



8.2 Is the result use in the experiment representative of the real result use patterns?
__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 3.3.*]
The result use is representative of the real result use in terms of, e.g.:
__type of use/search goals __effect on success criteria __effect on information needs
__other:_____ [__N __Y, to: UC 1.9.*, UC 3.3.*]
__the result use of end users is not known/well understood.
THE WHITEBOARD

9. EVALUATION CRITERIA [__Not applicable]
9.1 Are the success criteria in the experiment representative of end users' success criteria?
__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.9.*]
The success criteria are representative of end users success criteria in terms of, e.g.:
__volume of relevant results __time spent (if the goal is to spend time) __user satisfaction
__meeting user expectations __task completion __objectivity/subjectivity of criteria
__other:_____ [__N __Y, to: UC 1.9.*, UC 3.1.7]
__End users' success criteria not known/well-understood
__Evaluation is not based on user criteria, but:_____
9.2 Are the failure criteria in the experiment representative of end users' failure criteria?
__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.9.*]
The failure criteria are representative of end users' failure criteria in terms of, e.g.:
__time __effort __poor result quality __frustration __"out of queries" __other: _____ [__N __Y, to: UC 1.9.*]
__End users' failure criteria not known/well understood

Figure 8. Page 5 of the Form4 – evaluation features for the Enterprise search use case.

5 Discussion and conclusions, future work

The various use case domains and case studies given here have different starting points and varying practical constraints: some have been researched for decades, others are new; some have well-established business models with middlemen, stakeholders, information producers and content owners in a stable symbiosis, others again are in flux with a changing market arena. This means that the results given above are somewhat heterogeneous and while they do not provide immediate cook-book-like deployment instructions for other domains, their diversity is intended to serve as an inspiration for the approach of explicit user modeling used to guide evaluation.

References

[Amigó et al, 2012] Amigó, E., Corujo, A., Gonzalo, J., Meij, E., and de Rijke, M. Overview of RepLab 2012: evaluating online reputation management systems. In: Proc. CLEF 2012.

[Azzopardi 2011] Azzopardi, L. The Economics in Interactive Information Retrieval. In: Proc. of SIGIR 2011

[Border 2002] Broder, A. A taxonomy of web search. ACM SIGIR forum, 36(2), 2002.

[Cockburn 2002] Cockburn, A. Agile software development. Addison- Wesley, 2002.

[Constantine 2006] Constantine, L. and Lockwood, L. Software for use: A Practical guide to the models and methods of usage-centered design. Addison-Wesley, 2006.

[Garcia Seco Seco de Herrera et al, 2012] García Seco de Herrera, A. Tsikrika, T., Lupu, M., Petras, V., Gäde, M., Kleineberg, M., Choukri, K. Deliverable 4.2, Tutorial on Evaluation in the Wild, , PROMISE NoE, FP7 258191, 2012.

[Jacobson 1987] Jacobson, I. Object-oriented development in an industrial environment. In: Proc. of OOPSLA '87: Sigplan Notices, 22(12), 1987.

[Jacobson 1992] Jacobson, I., Christerson, M., Jonsson, P., and Overgaard, G. Object-Oriented Software Engineering: A Use Case Driven Approach. Addison-Wesley, 1992.

[Järvelin 2002] Järvelin, K. and J. Kekäläinen, J. Cumulated gain- based evaluation of IR techniques. ACM Trans. Inform. Syst. 20(4), 2002.

Järvelin et al, 2012 Järvelin, A., Eriksson, G., Hansen, P., Tsikrika, T., Garcia Seco de Herrera, A., Lupu, M., Gäde, M., Petras, V., Rietberger, S., Braschler, M., and Berendsen, R. Revised Specification of Evaluation Tasks, deliverable 2.2, PROMISE NoE, FP7 258191, February 2012.

[Keskustalo 2009] Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., and Lykke, M. 2009. Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case of Extremely Short Queries. In: Proc. of AIRS 2009.

Liu 2010 Liu, J., and Belkin, N. Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In: Proc. of SIGIR 2010.

[Moffat 2008] Moffat, A., and Zobel, J. Rank-Biased Preceision for Measurement of Retrieval Effectiveness. ACM Trans. Inform. Syst. 27(1), 2008.

[Murdock 2013] Murdock, V., Clarke, C., Kamps, J., and Karlgren, J. Proceedings of SEXI 2013 - Workshop on Search and Exploration of X-Rated Information at WSDM 2013.

[Müller et al, 2012] Müller, H., de Herrara, A. G. S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., & Eggel, I. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Proc. of CLEF, 2012.

[Rose 2004] Rose, D. and Levinson, D. Understanding user goals in Web search. In: Proc. of WWW'04, 2004.

[Sanderson, 2010] Sanderson, M. Test collection based evaluation of information retrieval systems. Now Publishers Inc, 2010.

Smucker 2012 Smucker, M. and Clarke, C. Time-based Calibration of Effectiveness Measures. In: Proc. of SIGIR 2012.

[Tsikrika et al, 2012] Tsikrika, T., Müller, H., Kahn Jr. C.E. Log analysis to understand medical professionals' image searching behaviour. In: Proc. of the 24th European Medical Informatics Conference. MIE2012, 2012.

[Wirfs-Brock 1993] Wirfs-Brock, R. Designing Scenarios: Making the Case for a Use Case Framework. Smalltalk Report, November-December, 1993.

6 Appendices

Appendix A. Background feature checklist

Background Feature Checklist – Form 1

1. USER ROLE	
1.1 Role name:	
1.2 Related roles:	
2. INCUMBENTS	**Not known/ not applicable** **Related to evaluation**
2.1 Domain knowledge	
__none __limited __moderate __high __varies	__N/K __N/A __N __Y:
2.2 General search or system proficiency	
__novice __moderate __expert __varies	__N/K __N/A __N __Y:
2.3 System knowledge	
__none __limited __moderate __high __varies	__N/K __N/A __N __Y:
2.4 Language skills	
__none __limited __moderate __high __varies	__N/K __N/A __N __Y:
2.5 Additional features (e.g. age, training, education, disabilities…)	
	__N/K __N/A __N __Y:
3. REPOSITORY	
3.1 Media	
__text __image __video __audio __graphs __3D __varies/other:	__N/K __N/A __N __Y:
3.2 Granularity	
__low __medium __high __varies __specific:	__N/K __N/A __N __Y:
3.3 Genre	
__commercial __factual __news __technical text __personal commentary __varies __other:	__N/K __N/A __N __Y:
3.4 Language	
__monolingual __bilingual __multilingual __other:	__N/K __N/A __N __Y:
3.5 Technical quality	
__low __moderate __high __varies	__N/K __N/A __N __Y:
3.6 Source dynamics	
__static __dynamic __stream __other:	__N/K __N/A __N __Y:
3.7 Indexing timeliness	
__immediate __every hour __daily __weekly __monthly __varies __other:	__N/K __N/A __N __Y:
3.8 Additional features/notes	
	__N/K __N/A __N __Y:
4. INFORMATION	
4.1 Origin of user input	
__aural __visual __mental __touch __varies __other:	__N/K __N/A __N __Y:
4.2 Clarity of information need	
__clear __medium __muddled __varies	__N/K __N/A __N __Y:
4.3 Flow direction	
__system to user __user to system __balanced __varies	__N/K __N/A __N __Y:
4.4 Information volume	
__low __medium __high __specific	__N/K __N/A __N __Y:
4.5 Complexity of information	
__low __medium __high __varies	__N/K __N/A __N __Y:
4.6 Additional features/notes	
	__N/K __N/A __N __Y:

5. INTERACTION	
5.1 Locus of control	
__push (system) __pull (user) __varies	__N/K __N/A __N __Y:
5.2 Complexity of interaction	
__low __medium __high __varies __specific:	__N/K __N/A __N __Y:
5.3 Predictability of interaction	
__low __medium __high __specific:	__N/K __N/A __N __Y:
5.4 Frequency	
__rare __recurrent __frequent __varied __specific:	__N/K __N/A __N __Y:
5.5 Regularity	
__irregular __regular period __varied __specific:	__N/K __N/A __N __Y:
5.6 Goal-orientation	
__random __vague __average __goal oriented	__N/K __N/A __N __Y:
5.7 Additional features/notes	
	__N/K __N/A __N __Y:
6. ORIENTATION	
6.1 Motivation	
__low __average __high __varies __specific:	__N/K __N/A __N __Y:
6.2 Likelihood of changing role	
__low __medium __high __specific: To what roles and when?	__N/K __N/A __N __Y:
6.3 Likelihood of abandoning system	
__low __medium __high __specific: On what conditions or why?	__N/K __N/A __N __Y:
6.4 Purpose of use	
__professional __leisure-utility __leisure-entertainment __other:	__N/K __N/A __N __Y:A
6.5 Optionality of use	
__required use __optional use (conditions):	__N/K __N/A __N __Y:
6.6 Additional features/notes	
7. RESTRICTIONS	
7.1 Cost of errors	
__low __medium __high __specific:	__N/K __N/A __N __Y:
7.2 Time restrictions	
__low __medium __high __specific:	__N/K __N/A __N __Y:
7.3 Access restrictions	
__none __pay-per-view __pay-per-search __pay-per-time __confidentiality/access rights __other:	__N/K __N/A __N __Y:
7.4 Device restrictions	
__size __input means __output means __processing speed __available tools or programs __other:	__N/K __N/A __N __Y:
7.5 Network restrictions	
__low __medium __high __varies __other:	__N/K __N/A __N __Y:
7.6 Additional restrictions related to organizational context (coverage etc.)	
	__N/K __N/A __N __Y:
7.7 Additional features/notes	
	__N/K __N/A __N __Y:

8. PHYSICAL ENVIRONMENT				
8.1 Mobility				
__mobile	__stationary	__varies	__specific:	__N/K __N/A __N __Y:
8.2 Geo-position				
__one	__many	__specific:		__N/K __N/A __N __Y:
8.3 Distractions				
__noise	__interruptions	__parallel tasks	__other:	__N/K __N/A __N __Y:
8.4 Climate and lighting conditions				
__lighting	__humidity	__temperature	__other:	__N/K __N/A __N __Y:
8.5 Additional features/notes				
				__N/K __N/A __N __Y:
9. SUCCESS CRITERIA				
9.1 Efficiency __				__N/K __N/A __N __Y:
9.2 Effectiveness __				__N/K __N/A __N __Y:
9.3 Satisfaction __				__N/K __N/A __N __Y:
9.4 System reliability __				__N/K __N/A __N __Y:
9.5 System intuitiveness __				__N/K __N/A __N __Y:
9.6 System comprehensibility __				__N/K __N/A __N __Y:
9.7 Actionability __				__N/K __N/A __N __Y:
9.8 Additional criteria __				__N/K __N/A __N __Y:
9.9 Notes				

Appendix B. Interaction checklist

Interaction and goals – Form 2

1. USE CASE NAME AND SUPPORTED USER ROLES	
1.1 Name:	
1.2 Supports user roles:	

2.USER GOALS	**Not known/ not applicable**	**Related to evaluation**
2.1 Type of information		
__single fact/answer/etc. __collection of facts/answers/etc. __single item (e.g. document) __collection of items __other:	__N/K __N/A	__N __Y:
2.2 Type of goal		
__viewing __exporting __navigating __manipulating __surfing __ordering/buying __other:	__N/K __N/A	__N __Y:
3. USE CASE RELATIONSHIPS		
3.1 Specializes:		
	__N/K __N/A	__N __Y:
3.2 Extends:		
	__N/K __N/A	__N __Y:
3.3 Uses:		
	__N/K __N/A	__N __Y:
3.4 Resembles:		
	__N/K __N/A	__N __Y:
4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE		

usecaseName	EXTENDS:
USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction	close (use case ends)
EXTENSIONS	

Appendix C. System and interface feature checklis

System and interface checklist – Form 3						
1. REQUEST FORMULATION					**Not known/ not applicable**	**Related to evaluation**
1.1 Supported search strategies						
__querying	__browsing	__monitoring	__other:		__N/K __N/A	__N __Y:
1.2 Query persistence						
__one shot	__permanent	__evolving	__other:		__N/K __N/A	__N __Y:
1.3 Query modality						
__text	__image	__video	__audio	__other:	__N/K __N/A	__N __Y:
1.4 Query formulation						
__specification	__example	__other:			__N/K __N/A	__N __Y:
1.5 Query language						
__simple keyword	__basic operators	__advanced	__specific:		__N/K __N/A	__N __Y:
1.6 Query target						
__content	__metadata/description	__other:			__N/K __N/A	__N __Y:
1.7 Query support						
__QE	__query suggestion	__translation	__spelling correction	__advanced query fields	__N/K __N/A	__N __Y:
__other:						
1.8 Browsing categories (content)						
__people	__country	__subject	__date/period	__media	__N/K __N/A	__N __Y:
__language	__collection	__other:				
1.9 Navigation support						
__sitemap	__FAQ	__classification	__thesauri	__other:	__N/K __N/A	__N __Y:
1.10 Changing between querying and browsing						
__supported	__not supported	__specific:			__N/K __N/A	__N __Y:
1.11 Additional features/notes						
					__N/K __N/A	__N __Y:
2. RESULT PRESENTATION						
2.1 Presentation hierarchy						
__one level	__two level	__other:			__N/K __N/A	__N __Y:
2.2 Presentation granularity						
__title	__summary	__metadata	__full item	__set of items	__N/K __N/A	__N __Y:
__other:						
2.3 Presentation organization						
__single item	__multiple items	__list	__ranked list	__browsing interface	__N/K __N/A	__N __Y:
__other:						
2.4 Result ordering						
__score	__date	__diversity	__author	__random	__N/K __N/A	__N __Y:
__other:						
2.5 Assessment support						
__scores	__highlighting	__popularity	__number of results	__relations within a doc.	__N/K __N/A	__N __Y:
__relations between docs.	__other:					
2.6 Additional features/notes						
					__N/K __N/A	__N __Y:
3. RESULT USE						
3.1 Manipulation						
__tagging	__annotation	__commenting	__discussing	__creating lists of documents	__N/K __N/A	__N __Y:
__other:						

3.2 On site consumption/use	
__viewing (on screen) __listening __analysis and interpretation __other:	__N/K __N/A __N __Y:
3.3 Exporting search context (queries, number of results etc.)	
__saving __printing __publishing __other:	__N/K __N/A __N __Y:
3.4 Exporting results	
__saving __printing __publishing __other:	__N/K __N/A __N __Y:
3.5 Sharing	
__exporting __within the system __other:	__N/K __N/A __N __Y:
3.6 Ordering/paying	
__internal __external __other:	__N/K __N/A __N __Y:
3.7 Additional features/notes	
	__N/K __N/A __N __Y:

Appendix D. Evaluation feature checklist

Evaluation Checklist – Form 4

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down under notes in the end of each section.

1. TEST SUBJECT CHARACTERISTICS [__Not applicable]	**RELEVANT U.C. FEATURES**
1.1 Are test persons representative of the end user population?	
__yes __reasonably __no __not known __other:	__None __feat._____
Test persons are representative of end user population in terms of:	
__demographics __search skills __language skills __domain knowledge __relation to search task __other_ __end user population not known/well understood	__None __feat._____
1.2 Notes	
	__None __feat._____
2. TOPICS [__Not applicable]	
2.1 Are topics representative of the real search topics/information needs?	
__yes __reasonably __no __not known __other:	__None __feat._____
Topics are representative of the real information needs in terms of:	
__domain of topics __type of search goal __clarity of information need __information need durability __other: __real information needs/search topics not known/well understood.	__None __feat._____
2.2 The topics are used for:	
__relevance __automatic runs __tasks given to test persons assessments __other:	__None __feat._____
2.3 Notes	
	__None __feat._____
3. REQUESTS [__Not applicable]	
3.1 Are requests representative of real requests?	
__yes __reasonably __no __not known __other:	__None __feat._____
Requests are representative of real requests in terms of:	
__type of request __request modality __query length __query quality __query structure __query durability __query formulation __other __real requests not known/well understood	__None __feat._____
3.2 Notes	
	__None __feat._____
4. DATA [__Not applicable]	
4.1 Is the data used in the evaluation activity representative of the real data?	
__yes __reasonably __no __not known __other:	__None __feat._____

The test data used is representative of the real data in terms of:	
__modality __dynamics __structure __intellectual content __size __curation __granularity __provenance __other: __real data not known/well understood	__None __feat._____
4.2 Notes	
	__None __feat._____
5. GROUND TRUTH CREATION [__Not applicable]	
5.1 Ground truth captures:	
__relevance of documents to topics __which of two ranked lists is better __ which of two documents is more relevant to a topic __other:	__None __feat._____
5.2 Ground truth is obtained:	
__manually __semi-automatically (e.g. pooling) __fully automatically __other:	__None __feat._____
5.3 Are relevance criteria representative of real users' relevance criteria?	
__yes __reasonably __no __not known __other:	__None __feat._____
Relevance criteria are representative of the real users' relevance criteria in terms of:	
__strictness of criteria __type of criteria __grades of relevance __other: __real users' relevance criteria not known/well understood	__None __feat._____
5.4 Are assessors representative of the end user population?	
__yes __reasonably __no __not known __other:	__None __feat._____
Assessors are representative of the end user population in terms of:	
__demographics __search skills __domain knowledge __language skills __relation to search task __other: __end user population not known/well understood	__None __feat._____
5.5 Are results shown to assessors representative of results shown to end users?	
__yes __reasonably __no __not known __other:	__None __feat._____
5.6 Notes	
	__None __feat._____
6. RESULT PRESENTATION [__Not applicable]	
6.1 Is the result presentation in experiment representative of target system(s)?	
__yes __reasonably __no __not known __other:	__None __feat._____
Result presentation is representative of target system(s) in terms of	
__presentation hierarchy __granularity __other: __target system result presentation not known	__None __feat._____
6.2 Notes	
	__None __feat._____

7. INTERACTION [__Not applicable]	
7.1 Interaction in the experiment is:	
__real user interaction __interaction model __other:	__None __feat._____
7.2 Is interaction in the experiment representative of real end user-system interaction?	
__yes __reasonably __no __not known __other:	__None __feat._____
The interaction is representative of real end user-system interaction in terms of:	
__search strategies __result assessment __goal orientation __learning __query formulation __session length/complexity __other: __real interaction patterns not known/well understood	__None __feat._____
7.3 Notes	
	__None __feat._____
8. RESULT USE [__Not applicable]	
8.1 Result use is included in evaluation with:	
__real users, real use __real users, controlled use __simulated __other	__None __feat._____
8.2 Is the result use in the experiment representative of the real result use patterns?	
__yes __reasonably __no __not known __other:	__None __feat._____
Result use is representative of the real result use in terms of:	
__type of use/search goals __effect on success criteria __effect on information needs __other: __ the result use of end users is not known/well understood	__None __feat._____
8.3 Notes	
	__None __feat._____
9. EVALUATION CRITERIA [__Not applicable]	
9.1 Are the evaluation criteria in the experiment representative of end users' success criteria?	
__yes __reasonably __no __not known __other:	__None __feat._____
The evaluation criteria are representative of end users success criteria in terms of:	
__volume of relevant results __time spent __user satisfaction __meeting user expectations __task completion __objectivity/subjectivity of criteria __other: __end users' success criteria not known/well understood __evaluation is not based on user criteria, but:	__None __feat._____
9.2 Are the evaluation criteria in the experiment representative of end users' failure criteria?	
__yes __reasonably __no __not known __other:	__None __feat._____
The evaluation criteria are representative of end users' failure criteria in terms of:	
__time __effort __frustration __poor result quality __"out of queries" __other: __end users' failure criteria not known/well understood	__None __feat._____

9.5 Notes	
	__None __feat._____
10. METRICS AND MEASUREMENTS	
10.1 Do the metrics measure what matters most to the end users?	
__yes __reasonably __no __not known __other:	__None __feat._____
Metrics measure what matters most to the end users in terms of:	
__task completion __cost of errors __efficiency __time spent __effort __domain restrictions __other: __what matters to end users not known/well understood	__None __feat._____
10.2 Are the metrics used predictive of real world performance?	
__yes __reasonably __no __not known __other:	__None __feat._____
The metrics are predictive of real world performance, in terms of:	
__relative performance between systems __absolute performance __other	__None __feat._____
10.3 Notes	
	__None __feat._____

Appendix E. Guidelines for the checklists

A use case based evaluation method – Documentation

This documentation aims to explain the four forms related to the use case based evaluation method developed within the course of the PROMISE project, in their current versions. The forms are intended to support a certain way of thinking about, designing, and documenting information access system evaluation. They are meant to be used as checklists for things to keep in mind while designing an evaluation. The forms are not questionnaires. Several example features given are relative, and the alternative answers given for the features are not meant to be precise, unambiguous scales, but rather examples of possible dimensions of the features. The forms are neither (and are not meant to be) exhaustive when it comes to the features and examples described in them: they present a selection of central features which may serve as a starting point when defining use cases for information access evaluation. Additional features may need to be described for a use case. Features that do not distinguish any role or that are unlikely to have significance for evaluation may be skipped.

Form 1: Background feature checklist

Form 1 aims at defining the operational context of a system that is useful for understanding the tasks and goals of the users, what kind of functional support they need from a system and what success criteria they have. It describes a set of needs, interests, expectations, behaviors and responsibilities characterizing a relationship between a class or a kind of users, i.e. a user role, and a system. Together with form 2 (which describes the tasks of the users in a certain role) it defines the problem space that is targeted in an evaluation.

The form is structured into groups of features illustrating different aspects of the operational context of a system. For each feature group a few example features are given. The features may apply to individual use cases, or to complete applications, or domains, and may therefore be defined on any suitable level of detail or generality. Use case domains (such as the PROMISE use case domains) may have specific features that all applications and use cases in that domain will inherit; An application will have some features that are shared by all use cases related to that application; and each individual use case may or may not diverge from other use cases related to similar applications or domains.

For this round of PROMISE use case work we suggest the following:

The focus of deliverable D2.4 should be on describing concrete use cases, not the generic and abstract "domain supercases" described in earlier deliverables. This means that the background features should be described for these concrete "subcases" of the domain supercases. However, as long as the supercases are not described separately, the features inherited from a supercase need to be described for each subcase.

It seems reasonable for each use case domain to focus on a single concrete use case and to define one set of background features for that use case. The background features should define the complete context of the system usage, including the domain related features and the use case specific features. It would be useful to keep track of the "origin" of the features, or the variation points, where the specific use case diverges (or specializes) from the domain supercase. If several subcases related to a supercase will be described, it clearly makes sense to first describe the general domain supercase, and then just describe the variation points for each of the subcases.

1. USER ROLE

1.1. User role name: A user role is an abstraction of a kind of users who have a particular relationship to a system. It is not a real person, job title, or a group of people. One role may be played by many users and one user may have many roles. The entire form 1 describes user role features, and this is where all that information is crystalized into a name of the role. Defining the user role for a specific use case, and not for the domain supercase, makes it concrete.

1.2 Related roles: 1.1 names the user role for one concrete subcase of the supercase in question; here the user roles for other subcases of that supercase, and the supercase itself might be listed**.**

2. INCUMBENTS

Incumbent is a term used for describing individuals holding a specific role. Each role is accompanied by certain expectations about the characteristics of the role incumbents. Thus the incumbent features represent the various bits of information about the actual users who are likely to play a particular role in relation to a system. We have listed some possible features below, but others may also be defined:

2.1 Domain knowledge: How much do incumbents know about the domain that the system supports.

2.2 General search/system proficiency: Refers to level of skill, or experience, in operation of the system.

2.3 System knowledge: refers to the (theoretical) knowledge of the incumbents concerning the specific system, how it operates and how to use it.

2.4 Language skills: A relative feature that refers to how well the incumbents command the language(s) needed for efficiently operating the system and carrying out their tasks. Depends on the data, task, domain, system, etc. Number of languages understood, level of reading and writing skills of the users, domain specific and technical language knowledge, or slang, dialect or historical language variant knowledge.

2.5 Other relevant user features: any potentially relevant information about the training, education, intelligence, sophistication or other socio-demographics of the incumbents, not elsewhere described.
2.6 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

3. REPOSITORY
This set of features describes the characteristics of the repository contents: the stored (or streamed) information items, the data.
3.1 Media: What modalities are present in the repository as the main content carriers, or as searchable content (such as text as means of accessing visual information)? Many repositories contain "multimodal" information items, such as text documents illustrated with images, or films subtitled with text, but the illustrations and the subtitles (might or) might not be independent carriers of information that would be searched for their own sake, and (might or) might not be directly accessible through the search system.
3.2 Granularity (of information items): What is the basic unit of information items in the repository? What is the granularity of the searchable content? Can users access the information down to specific data elements, paragraphs or film scenes? Or is retrieval limited to higher levels of granularity such as complete articles, books or films, or even newspapers, volumes, collections of films, etc.?
3.3 Genre: What's the genre of the information items?
3.4 Language: What languages or sublanguages and what type or level of language is present in the repository? It could be monolingual, multilingual, dialectal, historical, professional sublanguages, scientific, easy-reading etc.
3.5 Technical quality: Refers to the technical quality of the information items in the repository: image resolution, OCR quality, document formatting etc. Technical quality is mainly interesting as a restriction which might affect both the user interaction and the performance of the retrieval algorithm.
3.6 Source dynamics: Describes whether the repository is a collection or a stream and the change rate of the collection/stream.
3.7 Indexing timeliness: Reasonable indexing intervals are clearly related to the source dynamics – in a static collection, indexing is always timely.
3.8 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

4. INFORMATION
This set of features describes the nature of information. It includes the information flow manipulated by the incumbent and exchanged between incumbent and system: where the information originates and how it flows between user and system.
4.1 Origin of user input: where does the input to the system from the user in this role originate? What is the source of the immediate information need/request? Does the user come to think of it or receive it as an assignment from an external source? Is the origin aural, visual, or mental? Is it related to the user's own information need or task, or is the user acting as an information intermediary?
4.2 Clarity of information need: Are information needs typically well-defined and clear to the users, or complex, or muddled. This feature might relate to the complexity of the information, but also to the task stage or domain knowledge of the user.
4.3 Flow direction: Is information predominantly acquired from the user or provided to the user? (e.g., should emphasis be on clear presentation of information or ease of input?) Flow direction is closely related to feature 5.1 "locus of control": often these too features get values that are mirror image of each other.
4.4 Information volume: how much information is available (in the repository!), and is exchanged between the user and the system. How much information does the user need to input, or needs to be presented to the user?
4.5 Complexity of information: how complex is the information communicated between the system and the user? How complex is the information needed by the user and how complex is the information (in the repository) which needs to be presented to the user?
4.6 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

5. INTERACTION

The interaction features describe typical or expected patterns of system usage associated with a given role, including things such as frequency and periodicity of interaction. The characteristics of the interaction pattern may affect what system and interface features are central for the user (for ease and efficiency of use etc.) and thus the success criteria.

5.1 Locus of control (push-pull): is the interaction initiated and/or driven by the system or by the user? This feature is closely related to feature 4.3 (flow direction). Often the two features are mirror images of each other.

5.2 Complexity of interaction: How complex are the interactions carried out in this specific user role?

5.3 Predictability of interaction: Are interactions within this role predictable or variable?

5.4 Frequency: How often will the user take on this role? (How frequent is this interaction for a typical user?)

5.5 Regularity: is the usage regular, or more or less sporadic?

5.6 Goal-orientation of the interaction: is the user focused on working toward a well-defined goal, or does the interaction include a feature of more random surfing and information encountering?

5.7 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

6. ORIENTATION

This set of features describes the orientation or attitude that users in the specific role have towards a system (usage), towards the role they are holding and towards the task they are carrying out.

6.1 Motivation: How motivated are users (in the considered user role) to carry out the interaction, to reach their goals?

6.2 Likelihood of changing role: How likely are users (in the considered user role) to change to other user roles (such as from a curious content viewer to a customer)? To which roles and under which conditions?

6.3 Likelihood of abandoning system: How likely are users (in the considered user role) to give up using the system without having completed the interaction and reached their goal? Under which circumstances might that happen?

6.4 Purpose of use: What type of a task are users (in the considered user role) occupied with? Is it professional or leisurely; or utility or fun related?

6.5 Optionality of use: Is the system usage optional for the users (in the considered user role), or is it required? This is closely related to the likelihood of abandoning the system: if a user must carry out a task, and only can do it using the specific system, the user is less likely to give up the interaction in the first place.

6.6 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

7. RESTRICTIONS

These features describe the different restrictions which limit the interaction for a user role.

7.1 Cost of errors: What is at stake if the user and the system fail in completing the task correctly? What type and level of risk is associated with the interaction in the user role? Both end user and stakeholder costs can be described.

7.2 Time restrictions: Refer to any context factors that limit the length of the interaction, e.g., related to the urgency of the task and information need, but should not be factors related to the repository business model (as that is covered below).

7.3 Access restrictions: Refer to any designed restrictions to accessing the repository contents which restrict the interaction for the user role. They can be related to confidentiality and user access rights, or to the business model of the service provider etc., however they should be noticeable for the users (as direct access restrictions, costs, time limits, or more indirectly as rules and regulations of use).

7.4 Device restrictions: Describe the limitations of the physical equipment through which an incumbent interacts with the system. This includes the type of device (e.g. PC, smartphone, game console, e-book device) and the types of input and output means used (keyboard, keypad, touch pad, mouse, microphone, camera; screen (size limitations?), speakers, earplugs, paper, punch cards, clay tablets, …).

7.5 Network restrictions (latency/cost): Any restrictions related to the network connection and data traffic, may be related to data traffic speed or cost.
7.6 Additional restrictions related to organizational context (e.g., coverage requirements, etc.): Any other domain, organization or work task related restrictions that are not described by the other features.
7.7 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

8. PHYSICAL ENVIRONMENT
This set of features describes factors of the physical environment in which a user interacts with the system.
8.1 Mobility: Refers to issues related to users being on the move. Simply the fact that a user is on the move has consequences for the interaction, but there might also be other interesting issues. For example, is the user task related to the movement, or are there clearly identifiable patterns of movement?
8.2 Geo-position: Where in the world, or in what part of a city is the user, etc.?
8.3 Distractions: Refers to anything that distracts the user interacting with the system, such as noisiness, interruptions, handling multiple parallel tasks, etc.
8.4 Climate and lighting conditions: lighting, humidity, temperature, indoors/outdoors, etc.
8.5 Additional features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

9. SUCCESS CRITERIA
These features are for describing the central success criteria for a user role. The goal is to describe what makes a user satisfied. (Evaluation criteria is described in form 4.)
9.1 Efficiency: A system that helps users complete their tasks with minimum waste, time or effort is efficient (as defined in Kelly 2009). Covers both efficiency of use (flexibility of operation, adaption to level of user skills, clarity of presentation of information etc.) and system responsiveness.
9.2 Effectiveness: Relates to the accuracy and completeness of the retrieval results. How important are accuracy and completeness? Which is more important, or are they as important?
9.3 Satisfaction: Refers to a user's subjective satisfaction with the experience of using a system.
9.4 System reliability: Refers to both internal system reliability, and to reducing errors made by system users.
9.5 System intuitiveness: How intuitive is the system? How easy is it for a user to learn to use the system, and once learned, how easy is it for a user to remember how it is used? (Learnability and rememberability of the system.)
9.6 System comprehensibility (transparency): Can user understand how the system works, and why certain results are retrieved?
9.7 Actionability: of the information retrieved – usefulness of information for supporting some activity, actions or goal. Does the information enable the user to take action, and e.g., make a decision, or change role from a content viewer to a customer?
9.8 Additional criteria/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

Form 2: Interaction and goals – the (core of the) use case

A *use case* is description of one kind of use to which a system can be put. It describes a system usage that is complete and meaningful to a user. Form 2 is used for describing the central aspects of an evaluation use case, including relations to other use cases, goals related to the user role and typical interaction patterns between a user in that role and the system.

1. USE CASE NAME AND SUPPORTED USER ROLES

The name of the use case, and the supported user roles are not central in small scale exercises, where just one use case and one user role are considered. They become more useful when a larger collection of use cases and user roles need to be handled. In that case "supported user roles" may be used for writing down all user roles within and outside a use case domain which are supported by the use case.

2. USER GOALS

2.1 Type of information: Refers to what kind of information needs are typical for the user role: are users looking for single items or several items; for ready answers, facts or notifications, or for complete information items (such as documents) from which the information can be extracted by the users.

2.2 Type of goal: Refers to how the user intends to use the search result, and how much the intended use needs to be supported within the system. The type of goal affects the interaction pattern, and potentially the preferred result presentation.

3. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

The use case narrative describes the interaction between the user (in the role) and the system from an external, black-box view. The focus is on identifying the interaction points between the user and the system, and on thus defining the required functionality of the system. Use case narratives can follow a more or less structured format. In form 2, we present a structured form for writing use case narratives (following Wirfs-Brock, and C&L p. 101) where the narrative is divided into two parts, *user intentions* and *system responsibilities* that show what the user aims to do in each step of the interaction and what system responsibilities relate to each of the user intentions (see figure 1 for an example).

findingIllustration

USER INTENTION	SYSTEM RESPONSIBILITY
request illustration	
	show appropriate images
select image	
	show preview
confirm	
	insert image
	close
EXTENSIONS: browsingResults	

Figure 1

Use case narratives can be written on different levels of abstraction: *conventional use* cases describe rather concrete interactions and may make many assumptions concerning the systems and the user interface (*how* users can formulate requests, *how* systems should respond to specific user actions), while e*ssential use cases* strive towards a more abstract description of the user intentions and system responsibilities (*what* the user aims to do in each interaction step and *what* responsibilities the system has towards those aims, given user expectations). From evaluation point of view, it is more important to describe the "what" than the "how". Therefore we adopt the more abstract essential use cases that better support identifying the essential points of interaction between a user and a system, separating the real user intentions and goals from the currently prevalent interface designs and information access system implementations. This is not to say that each evaluation should come up with innovative interface and system design, but to stress that none of the interface and system designs should be taken as givens, but as choices which are made and which will affect the evaluation and the success of the implemented systems. The more general and abstract level of essential use cases also

makes it easier to recognize general similarities between use cases, which may help in categorizing use cases into groups where similar evaluation approaches may be useful.

As an example, the structured, essential use case narrative for a situation where a user needs to find an illustrative image for a newspaper article from an image archive is outlined in figure 1 above.

4. USE CASE RELATIONSHIPS

Use cases do not exist in isolation. A complete software system may have to support dozens or hundreds of interrelated use cases. Capturing the functional requirements of the complete system then requires – in addition to describing the individual use cases – also describing the relationships between the use cases. One use case could be a part of many higher level use cases, use cases might be composed of other use cases, or similar use cases could relate to different user goals served by the system. Careful definition of the use case relationships helps avoiding unnecessary work writing abundant use cases, and making use case models simpler. Below, the use case relationships specialization, extension, composition and affinity are discussed.

4.1 Specialization

Specialization is used for subtyping. It is a hierarchical "is-a" relationship between use cases, where a more general use case (*supercase*) has one or several specialized sub use cases (*subcases*). The supercase is then typically an abstract class of interaction, and the subcases its concrete manifestations, for example (example from C&L 2005):

withdrawingCash and **queryingStatus** are specializations of the supercase **usingATM**.

The supercase does not stand on its own as an interaction which would actually be carried out by a real user. However it can be used to describe the use case features that all the subcases will inherit. Thus specialization simplifies the overall use case model by allowing reuse of more general interaction patterns and use case features in the subcases without repeating them.

- The current PROMISE use cases are typically described on a supercase level. Based on the feedback from the use case validation interviews, the separate subcases are not properly identified, which makes the use cases too generic (and thus difficult to validate or disprove). Therefore, we suggest that while the background features may be defined to include both the supercase and the subcase (to define the domain), the subcases need to be clearly identified in the interaction model. If necessary, the subcase interaction models may point back to diversion points in the background features. Evaluation may focus on one subcase, or aim to cover all subcases of a supercase, but it is not evident that exactly the same evaluation approach or success metrics can be used in both situations.

4.2 Extension

Extension is used for describing optional interactions in a use case. It is a use case relationship that helps to keep a use case's main flow of interaction simple. One use case extends another use case if it represents inserted or alternative patterns of interaction to the extended use case's main flow. For example, in the previous use case for finding an illustrative image for a newspaper article, browsing through (several pages of) results may sometimes be needed, but it might not be included in the main flow of interaction, but instead handled as an extension use case **browingResults** (or browsingImages?) instead:

browsingResults	**EXTENDS: findingIllustration**
USER INTENTION	SYSTEM RESPONSIBILITY
Request more [continue until found] Select image	Show more images Close (return to main flow)

This way the main flow is kept simple. The extension does not appear in the narrative of the main flow and could occur at any point in the flow of interaction. This practice has the benefit of the interaction described in the extension becoming available also for other use cases. This way use cases can be written reusing other use cases as components.

- Goal: a limited set of reusable modules which includes a small number of common variations of simple main flows and a set of extensions that can be used for describing the pattern of interaction for most information access use cases. The first set of modules will be extracted from the use cases described by the PROMISE partners and presented in D2.4.

4.3 Composition

Composition describes required parts or subsequences of use cases. It is used for modeling how use cases are composed of subcases representing subordinate or included patterns of interaction; the interaction described in a supercase is carried out by making use of the interactions within the subcase(s). The subcases are required parts of the main flow and are carried out in the order described in the supercase narrative. Therefore, the narrative of the supercase will refer to all of the subcases included (unlike in *extension*, where the extensions are not visible to the use cases they extend and may occur at any point of the interaction in the main flow). For example, a retrieval system for accessing hospital patient records might use a subcase **authorizingAccess** for checking user access rights as a compulsory first step in the flow of interaction.

4.4 Affinity

Affinity represents apparent but unspecified relatedness between use cases. It can be used for grouping use cases to form meaningful clusters, even if the exact nature of the similarities between the use cases is unclear (e.g. using the ATM example, transferringFunds and depositioningFunds might be more closely related to each other than to withdrawingCash). On occasion, two use cases that represent different user intentions may be virtually identical when it comes to modeling the interaction and evaluation criteria for tasks they represent. (C&L: 114)

Form 3: System and interface feature checklist

Form 3 provides a checklist for describing the system and interface features which are central for defining evaluation tasks/experiments. The form should not be used as a "wish list" or a tool for sketching ideal user interface functionality, but instead be used for documenting the interface decisions made in the experimental design.

Even Cranfield style laboratory studies, where no users are involved, make assumptions concerning the user interface, e.g., concerning the request formulation and result presentation functionality: requests are often assumed to be formulated as unstructured keyword queries, and results presented as ranked lists of document ID's, titles or the like. Such assumptions have a major effect on the evaluation setup and the results and may thus limit or improve the realism and applicability of the results.

In typical information access system evaluations the system and interface features need to be defined for at least two central points of interaction: request formulation and result presentation. Other functionalities may also need to be considered, e.g., related to information use. Information use is included in form 3 ("result use"), but even other functionality (currently not covered by form 3) related to the central points of interaction identified in form 2 may need to be described.

The description is still black-box: only the interface features visible to the user are described. Just as with forms 1 and 2, the feature categories and the example features are intended as helpful examples. Some of them may be skipped and additional features may be considered.

1. REQUEST FORMULATION

The request formulation features describe the request formulation functionality of the evaluated system. In the simplest case, only short keyword queries might be supported.

1.1 Supported search strategies: Describes which general approaches to searching are supported. The most obvious alternatives are querying, browsing through predefined content categories, and monitoring.

1.2 Query persistence: Especially monitoring and filtering queries can be permanent or slowly evolving, but also other types of queries can be recurring.

1.3 Query modality: What query modality or modalities are supported? Text, visual, audio, etc.

1.4 Query formulation: Are queries specified by the users or can example documents be used as queries (typing or drawing a query as opposed to using an example document or image as a query).

1.5 Query language: What query language(s) are supported? How complex or advanced operators are supported?

1.6 Query target: What is searchable: full content of documents, or metadata? (cf. feature 3.2 "granularity" in form 1.)

1.7 Query support: What kind of support is offered to the user (automatically or prompted by user)? Is support offered for handling inflection, spelling, tokenization or OCR errors; is support offered for formulating advanced queries, or for translation or coming up with alternative query terms?

1.8 Browsing (content) categories: How is the content of the repository made accessible using browsing? How is the content classified or categorized for browsing?

1.9 Navigation support: Is navigation supported and how?

1.10 Changing between querying and browsing: Is changing between querying and browsing supported, and how?

1.11 Additional query formulation features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

2. RESULT PRESENTATION

This set of features present the result presentation functionality of the evaluated system.

2.1 Presentation hierarchy: Result presentation is often divided into a hierarchy where a large number of condensed results typically are presented first. The number of presented items gets smaller and the level of detail higher for the following levels of hierarchy. Many information retrieval studies consider only a one-level result presentation hierarchy, where only ranked lists of document pointers

are considered, while operational systems very often divide the result presentation into at least two levels: ranked lists of document titles (and summaries) on the first level, and full documents on the second level.

2.2 Presentation granularity: What is shown to the user: a complete information item, part of an item, metadata related to an item, aggregation of items?

2.3 Presentation organization: How are the results presented to the user? Does the user see one result at a time, or an ordered list of items, or a grid of thumbnails, etc.?

2.4 Result ordering (by): Based on what criteria is the result ordered and can the user reorder the result by different criteria?

2.5 Assessment support: What kind of support is offered for making it easier to identify relevant results from result lists, and for assessing the relevancy of documents?

2.6 Additional result presentation features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

3. RESULT USE

The result presentation features can be used for describing how result use is supported in the evaluated system.

3.1 Manipulation: Can the user manipulate or change the information items within the system, or participate in content production? Can the user e.g., modify the information items, or comment, annotate or discuss them?

3.2 On site consumption/use: What kind of result use within the system is supported? Can full documents be viewed, played, etc. within the system (on screen, …)? Are there tools for improving readability, visualizations for supporting interpretation, making notes? Is support for visually impaired provided, etc.

3.3 Exporting search context: Is exporting the search context supported?

3.4 Exporting results [single items/sets of documents]: Is exporting results supported?

3.5 Sharing: Is sharing results with collaborators (within or outside of the system) or in social media supported?

3.6 Ordering/paying: Is ordering/paying (or illegal downloading) supported within the system, or externally?

3.7 Additional result use features/notes: This field occurs at the end of every feature group and can be used for making notes concerning additional features or other thoughts related to the feature group.

Form 4: Evaluation

We envision that most commonly, when somebody picks up these four use case forms, she already has a specific evaluation task in mind. The forms will then be used for documenting the task, and as a help for making the experiment more realistic and correct, given the use case underlying the task. The three first forms document the use case, the real thing. The fourth one (the one discussed now) documents the relation between the use case and an experiment and is thus aimed for supporting experimental design based on use cases. This means that the fourth form is the most central one, and also typically requires more thought than the other three.

Form 4 is divided into 10 sections each describing a factor, or component prevalent in experimental design in information access studies. These components reflect to some extent the TREC-style laboratory experiment model. However, the aim is to cover a much larger spectrum of evaluation approaches. Therefore, each section should be understood in a more general manner than the definition in a TREC experiment might be. Each section is shortly described below.

The sections typically follow the same format of asking how representative each evaluation component is of the real world situation, and then asking to specify in what way it is, or is not, representative of the real world situation. The alternatives given for the follow-up questions are not necessarily exhaustive, but are aimed as examples of potentially important aspects. Considering (and documenting) additional aspects is encouraged.

The second column is for noting the connections to the use case features defined in forms 1-3. To connect the use case features to the different components of the experiment, list for each (relevant) component the relevant use case features. For example, incumbent features are likely to be relevant for defining desirable test subject or relevance assessor characteristics; interaction features are likely to be relevant for designing the interaction component. However, many less obvious connections may exist for most of the evaluation components.

After each section, there is a “whiteboard”, a space for writing notes. This is where you can note additional aspects of the evaluation component that should be considered, as well as make notes considering the detected divergences between the experiment and the use case. Noting the divergences is essential for identifying the weaknesses of the intended experimental design. Many of the issues may be fixable with minor re-thinking of the experiment; other may be more difficult or even impossible to avoid, but need to be considered when analysing results and their applicability.

1. TEST SUBJECT CHARACTERISTICS
If test subjects are used in an experiment, then it should be considered how they reflect the characteristics of the intended end users of the system (described in the use case): do they have the same demographic background, similar knowledge and skills with respect the search task an operating the system, are they likely to interact with the system in a similar way. If the test subjects are not recruited from the end user population, it should be considered how the (potential) differences will affect the evaluation. If there are no test subjects, choose “not applicable”.

2. TOPICS
Many experiments make use of pre-defined search “topics” or descriptions of information needs. The exact content or format of a topic depends on the domain of the evaluation and on what the topics are used for: are they intended to serve as, e.g. descriptions of search tasks given to test persons, used as definitions of relevance criteria in the relevance assessment process, or as a source of query words for automatic runs. When an evaluation is not based on real users interacting with a system to solve their own search tasks, the topics easily begin diverging from the real information needs, e.g., due to lack of real interaction with and learning from the retrieved information, which may make the topics easily undesirably static. Factors related to topics that could be considered, include e.g.:

Domain: Do the topics reflect the typical domain of the information needs, whatever the typical domain? (e.g., "expert biomedical" and "general Web" could both be domains)

Clarity of information need: Do the topics reflect the level of clarity or vagueness typical of the real information needs?

Type of search goal: Do the topics reflect the typical search goals, as defined in Form 2.

Durability of information needs: This feature refers to how static or dynamic information needs are within search session, but also in longer perspective: if information needs tend to be recurrent (or continuous), or only to occur once.

3. REQUESTS

"Requests" refers to all kinds of expressions of information needs, formulated in a form understandable to an information access system. They can be either momentary/ad hoc queries, or to a varying degree static queries, such as filtering, monitoring, profiling or routing queries, or clicks on predefined "browsing categories". In other words, in a typical study, requests are the concrete expressions of the information needs (described in the topics) that the "search engine" gets to work on.

4. DATA

Information access systems might be evaluated either using the real data, or some sort of "surrogate" data consisting of a part of the real data, or similar data streams or collections. When evaluation is based on surrogate document collections, issues such as type of data or documents included in the test collection, size of the test collection and the test collection dynamics become critical for the validity and applicability of evaluation.

5. GROUND TRUTH CREATION

Ground truth refers to the criteria by which an information access system is evaluated. In information retrieval evaluation practice, ground truth typically describes relations between information needs and documents. The most central relation is the relevance of the documents, given an information need. Even other things can be included in the ground truth, such as the target audience of the documents.

5.1 Ground truth captures

What does the ground truth aim to capture?

5.2 Ground truth (typically relevance assessments) is obtained

Ground truth can be obtained either through manually judging the relevance of a set of documents for each information need, or by different (more or less) automated means. How is the ground truth obtained, and how much manual work or human judgment is involved? How does this affect the coverage and reliability of the ground truth?

5.3 Are the relevance criteria used representative of the real users' relevance criteria?

The question and the alternatives seem obvious.

5.4 Are assessors representative of the end user population?

The question itself and most of the alternatives seem obvious.

Relation to search task refers to the personal relation: end users may or may not be using information access systems to satisfy their own information needs (informaticists typically try to satisfy information needs of others…). Assessors may have a similar relation to the search task, but are most often working on artificial search tasks that they do not have any personal relation to.

6. RESULT PRESENTATION

In some studies, especially interactive user studies and interaction simulations, the role of result presentation is obvious, and may even be the focus of the studies. But even in many other types of studies, result presentation needs to be considered. E.g., Cranfield-style laboratory studies typically base the evaluation on ranked lists of results, while other options could be possible.

Hierarchy and granularity refer to issues discussed in Form 3:
Presentation hierarchy: Results are often presented in a hierarchy, where the first level typically contains condensed information about many items (e.g., ranked list of document titles), and more information concerning fewer documents is presented on the following levels (e.g., one full document).
Presentation granularity: What is shown to the user: a complete information item, part of an item, metadata related to an item, aggregation of items?

7. INTERACTION

Interaction can be included in information access studies in different ways: as natural interaction of real users with systems in observational or log studies, as interaction in more or less natural surroundings in controlled user studies, or as interaction simulations where real users are replaced by (statistical) interaction models. In some other studies, interaction is completely ruled out, which also affects the applicability of the results. Interesting aspects related to interaction include, but are not limited to:

Search strategies: Are the typical search strategies covered in the experiment? Querying, browsing and monitoring lead to different types of interaction patterns.
Goal orientation: How focused is the user working toward a goal, or how much of randomness is there in the interaction. Do users know where they are going and to they stay on the path?
Learning: Really closely connected to information need durability.
Session length/complexity: How long, or complex, are typical search sessions – could be defined e.g., as time, or as number of user actions.
Query reformulation (strategies, cost, ...): Do users tend to reformulate their queries, or are one-shot queries more common? What can lead to query reformulation: quality of search results, learning, changing information needs, etc.? What kind of strategies do users have for query reformulation (e.g., adding terms, replacing terms, using a completely new query)? Cost models for query reformulation (and other user actions) are typically defined in interaction simulations.

8. RESULT USE

Information access is rarely a goal in itself, and thus result use (or information use) is an important part of users' information seeking cycle. What are the most common result use patterns of users and how do they affect the search interaction, learning, and success criteria of the users?

9. EVALUATION CRITERIA

Typical evaluation criteria for information access studies focus on the quality of the search results. The selected evaluation criteria should reflect the success and failure criteria of the envisioned end users of the system and thus the variety of success criteria specified in Form 1 should be considered.

Moreover, end users are not always at the center of evaluation. In many cases, information access systems are support functions to some other services or products. The most important evaluation criteria may then relate to how well the system supports the business of the service provider: how it affects revenue, traffic, or conversion rate (of e.g., visitors to paying customers, or registered members). Evaluation against a strong "gold-standard" might not be interesting, if it's enough that a system allows the users to do what they aim to do well enough to keep them from going somewhere else with their business.

The aspect of "well-enough" or an acceptable level of performance is often neglected. Evaluation may be based on, and should at least be aware of the failure criteria of the envisioned end users.

Prioritizations, generalizations and simplifications over the real success criteria are often necessary for creating feasible and usable evaluation criteria. If the real criteria are then not well understood, it will be difficult to assess how the simplifications may affect the validity and applicability of the results.

10. METRICS AND MEASUREMENTS

Evaluation criteria need to be operationalized in some stable and feasible manner that makes comparisons between systems possible. Even the metrics should reflect the end users' (or whatever) success and failure criteria.

Appendix F. Visual clinical decision support checklists

Background Feature Checklist – Form 1

1. USER ROLE

1.1 Role Name CLINICIAN

1.2 Related Roles RADIOLOGIST

	Not known/ not applicable	Related to evaluation
2. INCUMBENTS		
2.1 Domain knowledge: X none __limited __moderate __high __varies	[__N/K __N/A]	[X N __Y, to:____]
2.2 General search or system proficiency: X novice __moderate __expert __varies	[__N/K __N/A]	[X N __Y, to:____]
2.3 System knowledge: X none __limited __moderate __high __varies	[__N/K __N/A]	[X N __Y, to:____]
2.4 Language skills: __ none __limited X moderate __high __varies	[__N/K __N/A]	[X N __Y, to:____]
2.5 Other relevant user features (e.g., age, training, education, disabilities, etc.): ____	[__N/K X N/A]	[X N __Y, to:____]
2.6 Additional features/notes: ____	[__N/K __N/A]	[__N __Y, to:____]
3. REPOSITORY		
3.1 Media: __text __image __video __ audio __ graphs __3D objects X varies __other: ____	[__N/K __N/A]	[__N X Y, to: DATA]
3.2 Granularity (of what is an information item): __low __ medium X high __varies __specific: articles	[__N/K __N/A]	[__N __Y, to: DATA]
3.3 Genre: __news __factual __entertainment X scientific __commercial __personal commentary __technical text __varies __other: ____	[__N/K __N/A]	[__N X Y, to: DATA]
3.4 Language: __monolingual __bilingual X multilingual __ other: ____	[__N/K __N/A]	[__N X Y, to: DATA]
3.5 Technical Quality: __low X moderate __high __varies	[__N/K __N/A]	[X N __Y, to:____]
3.6 Source Dynamics: __static collection X dynamic collection __stream __other: ____	[__N/K __N/A]	[X N __Y, to:____]
3.7 Indexing Timeliness: __immediate __every hour __daily __weekly __monthly X varies __other: ____	[__N/K __N/A]	[X N __Y, to:____]
3.8 Additional features/notes: ____	[__N/K __N/A]	[__N __Y, to:____]
4. INFORMATION		
4.1 Origin of user input: __aural __visual __mental __touch X varies __specific: ____	[__N/K __N/A]	[__N __Y, to:____]
4.2 Clarity of information need: __clear X medium __muddled __varies	[__N/K __N/A]	[__N __Y, to:____]
4.3 Flow direction: X system to user __user to system __balanced __varies	[__N/K __N/A]	[__N __Y, to:____]
4.4 Information volume: __low X medium __high __specific: ____	[__N/K __N/A]	[__N X Y, to: GROUND TRUTH]
4.5 Complexity of information: __low __medium X high __varies	[__N/K __N/A]	[__N X Y, to: DATA]
4.6 Additional features/notes: ____	[__N/K __N/A]	[__N __Y, to:____]

5. INTERACTION

5.1 Locus of control: __push (system) ✕pull (user) __varies [__N/K __N/A] [✕N __Y, to:____]

5.2 Complexity of interaction: ✕low __medium __high __varies __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

5.3 Predictability of interaction: __low ✕medium __high __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

5.4 Frequency: __rare __recurrent ✕frequent __varied __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

5.5 Regularity: __irregular ✕regular period __varied __specific:____ [__N/K __N/A] [✕N __Y, to:____]

5.6 Goal-orientation: __random __vague __average ✕goal oriented __other:____ [__N/K __N/A] [✕N __Y, to:____]

5.7 Additional features/notes: ____ [__N/K __N/A] [__N __Y, to:____]

6. ORIENTATION

6.1 Motivation: ✕high __middle __low __varies __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

6.2 Likelihood of changing role: __low ✕medium __high __specific: ____

To what roles and when? ____ [__N/K __N/A] [✕N __Y, to:____]

6.3 Likelihood of abandoning system: __low ✕medium __high __specific: ____

On what conditions or why? failed search [__N/K __N/A] [✕N __Y, to:____]

6.4 Purpose of use: ✕Professional __leisure/utility __leisure/entertainment __other: ____ [__N/K __N/A] [✕N __Y, to:____]

6.5 Optionality of use: __ Required use ✕optional use (conditions): ____ [__N/K __N/A] [✕N __Y, to:____]

6.6 Additional features/notes: ____ [__N/K __N/A] [__N __Y, to:____]

7. RESTRICTIONS

7.1 Cost of Errors: ✕low __medium __high __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

7.2 Time restrictions: __none ✕low __medium __high __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

7.3 Access restrictions: ✕none __confidentiality/access rights __pay-per-view __pay-per-search __pay-per-time __other:____ [__N/K __N/A] [__N ✕Y, to: DATA]

7.4 Device restrictions: __size ✕processing speed __available other tools or programs __input means _output means __other ____ [__N/K __N/A] [✕N __Y, to:____]

7.5 Network restrictions: __low __medium ✕high __varies __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

7.6 Additional restrictions and requirements related to organizational context (e.g., coverage requirements, etc.):

____ [__N/K __N/A] [✕N __Y, to:____]

7.7 Additional features/notes: ____ [__N/K __N/A] [__N __Y, to:____]

8. PHYSICAL ENVIRONMENT

8.1 Mobility: __mobile ✕stationary __varies __specific: ____ [__N/K __N/A] [✕N __Y, to:____]

8.2 Geo-position: __one ✓ many __specific:_____ [__N/K __N/A] [✓ N __Y, to:____]
8.3 Distractions: __noise __interruptions __parallel tasks __other:_____ [__N/K ✓ N/A] [✓ N __Y, to:____]
8.4 Climate and lighting conditions: __lighting __temperature __humidity __other:_____ [__N/K ✓ N/A] [✓ N __Y, to:____]
8.5 Additional features /notes:_____ [__N/K __N/A] [__N __Y, to:____]

9. SUCCESS CRITERIA

9.1 Efficiency: ✓ [__N/K __N/A] [✓ N __Y, to:____]
9.2 Effectiveness: ✓ [__N/K __N/A] [__N ✓ Y, to: EVALUATION CRITERIA]
9.3 Satisfaction: ___ [__N/K __N/A] [✓ N __Y, to:____]
9.4 System reliability: ✓ [__N/K __N/A] [✓ N __Y, to:____]
9.5 System intuitiveness: ___ [__N/K __N/A] [✓ N __Y, to:____]
9.6 System comprehensibility: ___ [__N/K __N/A] [✓ N __Y, to:____]
9.7 Actionability: ___ [__N/K __N/A] [✓ N __Y, to:____]
9.8 Additional criteria: ___ [__N/K __N/A] [__N __Y, to:____]
9.9 Notes on success criteria:

_____

_____

_____

Form 2 – Interaction and goals

1. USE CASE NAME AND SUPPORTED USER ROLES

1.1 Name: VISUAL CLINICAL DECISION SUPPORT FOR MEDICAL DIAGNOSIS

1.2 Supports (user roles): CLINICIANS

2. USER GOALS

Not Know/ Applicable | Related to Evaluation

2.1 Type of information: __single fact/answer/notification __collection of facts/answers/notifications __single item (e.g., document) X collection of items __other:_____ [__N/K __N/A] [__N X Y, to: DATA]

2.2 Type of goal: __viewing __exporting X navigating __ordering/buying (transactional) __manipulating __surfing __other:_____ [__N/K __N/A] [X N __Y, to:____]

3. USE CASE RELATIONSHIPS

3.1 Specializes: QUERYNG IMAGES, QUERYING DOCUMENTS [__N/K __N/A] [__N X Y, to: TOPICS

3.2 Extends: BROWING RESULTS [__N/K __N/A] [X N __Y, to:____]

3.3 Uses: TEXT RETRIEVAL, IMAGE RETRIEVAL [__N/K __N/A] [__N X Y, to: EVALUATION CRITERIA

3.4 Resembles: RETRIEVAL [__N/K __N/A] [__N X Y, to: "]

4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

usecaseName	EXTENDS:
USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction • FORMULATE THE QUERY (TEXT, IMAGES, STRUCTURED DATA) • PERUSES FIRST RESULT PAGE AND CLICKS ON FEW RESULTS • CLINICIAN FINDS THE IMAGES AND ARTICLES	• RETRIVE THE RESULTS ACCORDING TO THE DEFINED CRITERIA • EVERY TIME IS CLICKED, THE SYSTEM PRESENTS THE FULL ARTICLE FROM THE BIOMEDICAL LITERATURE • SUCCESS Close (Use Case Ends)
EXTENSIONS	

Form 3 - System and interface feature checklist

	Not Known/ Applicable	Related to Evaluation
1. REQUEST FORMULATION		
1.1 Supported search strategies:		
✗querying __browsing __monitoring __other:_____	[__N/K __N/A]	[__N ✗Y, to: EVALUATION CRITERIA]
1.2 Query persistence: __one shot ✗permanent __evolving __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
1.3 Query modality: ✗text ✗image __video __audio __other:_____	[__N/K __N/A]	[__N ✗Y, to: EVALUATION CRITERIA]
1.4 Query formulation: ✗specification __example __other:_____	[__N/K __N/A]	[__N ✗Y, to: 4]
1.5 Query language:		
__simple keyword __basic operators ✗advanced __specific:_____	[__N/K __N/A]	[__N ✗Y, to: 4]
1.6 Query target: ✗content __metadata/description __other:_____	[__N/K __N/A]	[__N ✗Y, to: 11]
1.7 Query support:		
__spelling correction __query suggestion __translation ✗advanced query fields (support for advanced query language __QE __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
1.8 Browsing (content) categories:		
__people _country __subject __media __date __period __language ✗collection __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
1.9 Navigation support:		
__sitemap __FAQ __classifications __thesauri __other:_____	[✗N/K __N/A]	[✗N __Y, to:_____]
1.10 Changing between querying and browsing:		
__supported ✗not supported __specific:_____	[__N/K __N/A]	[✗N __Y, to:_____]
1.11 Additional query formulation features/notes:_____	[__N/K __N/A]	[__N __Y, to:_____]
2. RESULT PRESENTATION		
2.1 Presentation hierarchy: ✗one level __two level __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
2.2 Presentation granularity:		
__title __summary __metadata ✗full information item __set of items __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
2.3 Presentation organization:		
__single item ✗multiple items __list __ranked list __browsing interface __other:_____	[__N/K __N/A]	[✗N __Y, to:_____]
2.4 Result ordering (by):		

X score __date __diversity __author __random __other:_____ [__N/K __N/A] [__N X Y, to: EVALUATION CRITERIA]

2.5 Assessment support:

__highlighting X scores __popularity __number of results __relations within a document

__relations between documents __other: _____ [__N/K __N/A] [__N X Y, to: " "]

2.6 Additional result presentation features/notes:_____ [__N/K __N/A] [__N __Y, to:_____]

3. RESULT USE

3.1 Manipulation:

__tagging __annotation __commenting __discussing __creating lists of documents __other:_____ [__N/K X N/A] [X N __Y, to:_____]

3.2 On site consumption/use: __viewing (on screen)__listening (within the system) __analysis and interpretation

__Other:_____ [__N/K X N/A] [X N __Y, to:_____]

3.3 Exporting search context [queries/number of results]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K X N/A] [X N __Y, to:_____]

3.4 Exporting results [single items/sets of documents]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K X N/A] [X N __Y, to:_____]

3.5 Sharing: __within the system __exporting __other:_____ [__N/K X N/A] [X N __Y, to:_____]

3.6 Ordering/paying: __internal X external __other: _____ [__N/K __N/A] [X N __Y, to:_____]

Evaluation – Form 4

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down on the "whiteboard" under each section.

1. TEST PERSON CHARACTERISTICS [X Not applicable] **RELEVANT U.C FEATURES**

1.1 Are test persons representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Test persons are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task (e.g. motivation)

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

THE WHITEBOARD

2. TOPICS [__Not applicable]

2.1 Are the topics representative of the real search topics/information needs?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Topics representative of the search topics/real information needs in terms of, e.g.:

X domain of topics __clarity of information need (clear/muddled): X type of search goal

__information need durability.

__other:_____ [__N __Y, to:_____]

__Real information needs/search topics not known/well understood

2.2 The topics are used for

X relevance assessments X automatic runs __tasks given to test persons

__other:_____ [__N __Y, to:_____]

THE WHITEBOARD

3. REQUESTS [__Not applicable]

3.1 Are the requests representative of the real requests?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The requests are representative of the real requests in terms of, e.g.:

__type of request __request modality __query length __query quality __query structure
__query formulation (specification/example) __query durability
__other:_____ [__N __Y, to:_____]
__Real requests not known

THE WHITEBOARD

4. DATA [__Not applicable]

4.1 Is the data used in the evaluation activity representative of the real data?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The test data used is representative of the real data in terms of e.g.:

__intellectual content __modality __size __dynamics __curation __provenance (created by one/created by many)
__structure __granularity
__other:_____ [__N __Y, to:_____]
__Real data not known [__N __Y, to:_____]

THE WHITEBOARD

5. GROUND TRUTH CREATION [__Not applicable]

5.1 Ground truth captures:

X relevance of documents to topics __which of two documents is more relevant to a topic

__which of two ranked lists is better __other:_____ [__N __Y, to:_____]

5.2 Ground truth is obtained:

__ manually X semi-automatically (e.g. pooling) __fully automatically __other:_____ [__N __Y, to: _____]

5.3 Are the relevance criteria representative of the real users' relevance criteria?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Relevance criteria are representative of the real users' relevance criteria in terms of e.g.:

X strictness of criteria __type of criteria (e.g. topicality or novelty) __grades of relevance

__other:_____ [__N __Y, to:_____]

__Real users' relevance criteria not known/well understood

5.4 Are assessors representative of the end user population?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Assessors are representative of the end user population in terms of, e.g.:

__demographics __search skills X domain knowledge __language skills __relation to search task

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

5.5 Are the results shown to assessors representative of the real results shown to end users?

__yes X reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

6. RESULT PRESENTATION [X Not applicable]

6.1 Is the result presentation in experiment representative of target system(s)?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result presentation is representative of target system(s) in terms of e.g.:

__presentation hierarchy __granularity __other:_____ [__N __Y, to: _____]

__Target system result presentation not known

THE WHITEBOARD

7. INTERACTION [X Not applicable]

7.1 Interaction in the experiment is:

__real user interaction __interaction model __other:_____ [__N __Y, to: _____]

7.2 Is the interaction in the experiment representative of real end user-system interaction?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The interaction is representative of real end user-system interaction in terms of e.g.:

__search strategies __result assessment __goal orientation __learning

__session length/complexity __query reformulation (strategies, cost, ...)

__other:_____ [__N __Y, to: _____]

__Real interaction patterns not known/well understood

THE WHITEBOARD

8. RESULT USE [X Not applicable]

8.1 Result use is included in evaluation with:

__real users, real use __ real users, controlled use __simulated __no result use_____ [__N __Y, to: _____]

8.2 Is the result use in the experiment representative of the real result use patterns?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result use is representative of the real result use in terms of, e.g.:

__type of use/search goals __effect on success criteria __effect on information needs

__other:_____ [__N __Y, to: _____]

__the result use of end users is not known/well understood.

THE WHITEBOARD

9. EVALUATION CRITERIA [__Not applicable]

9.1 Are the success criteria in the experiment representative of end users' success criteria?

Xyes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The success criteria are representative of end users success criteria in terms of, e.g.:

Xvolume of relevant results __time spent (if the goal is to spend time) __user satisfaction

__meeting user expectations __task completion __objectivity/subjectivity of criteria

__other:_____ [__N __Y, to: _____]

__End users' success criteria not known/well-understood

__Evaluation is not based on user criteria, but:_____

9.2 Are the failure criteria in the experiment representative of end users' failure criteria?

Xyes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The failure criteria are representative of end users' failure criteria in terms of, e.g.:

__time __effort Xpoor result quality __frustration __"out of queries" __other: _____ [__N __Y, to: _____]

__End users' failure criteria not known/well understood

10. METRICS [__Not applicable]

10.1 Do the metrics measure what matters most to the end users?

__yes __reasonably __no, not very __ that's not the goal __not known

__other:_____ [__N __Y, to: _____]

Metrics measure what matters most to the end users in terms of, e.g.:

__task completion __cost of errors __efficiency __time spent __effort __domain restrictions

__other:_____ [__N __Y, to: _____]

10.2 Are the metrics used predictive of real world performance?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The metrics are predictive of real world performance, in terms of, e.g.:

__relative performance (between systems) __absolute performance __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

Appendix G. Intellectual property checklists

Background Feature Checklist – Form 1

1. USER ROLE

1.1 Role Name_ **Searcher** _____

1.2 Related Roles_ **none** _____

	Not known/ not applicable	Related to evaluation
2. INCUMBENTS		
2.1 Domain knowledge: __ none __limited __moderate __ high __varies	[__N/K __N/A]	[__N __Y, to: 2.1]
2.2 General search or system proficiency: __ novice __moderate __ expert __varies	[__N/K __N/A]	[__N __Y, to:_____]
2.3 System knowledge: __ none __limited __moderate __high __varies	[__N/K __N/A]	[__N __Y, to:_____]
2.4 Language skills: __ none __limited __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: 2.1]
2.5 Other relevant user features (e.g., age, training, education, disabilities, etc.): _most **searchers** have at least BSc_____	[__N/K __N/A]	[__N __Y, to:_____]
2.6 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]
3. REPOSITORY		
3.1 Media: __text __image __video __ audio __ graphs __3D objects __ varies __other: _formulae_____	[__N/K __N/A]	[__N __Y, to: 4]
3.2 Granularity (of what is an information item): __low __ medium __high __varies __specific: _____	[__N/K __N/A]	[__N __Y, to: 4]
3.3 Genre: __news __factual __entertainment __scientific __commercial __personal commentary __technical text __varies __other: _ legal_____	[__N/K __N/A]	[__N __Y, to: 4]
3.4 Language: __monolingual __bilingual __multilingual __ other: _____	[__N/K __N/A]	[__N __Y, to: 4]
3.5 Technical Quality: __low __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: 4]
3.6 Source Dynamics: __static collection __dynamic collection __stream __other: __**depends (varies)** _____	[__N/K __N/A]	[__N __Y, to: 4]
3.7 Indexing Timeliness: __immediate __every hour __daily __weekly __monthly __varies __other: _____	[__N/K __N/A]	[__N __Y, to:_____]
3.8 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]
4. INFORMATION		
4.1 Origin of user input: __aural __visual __mental __touch __varies _specific:_**information need is assigned to searcher**	[__N/K __N/A]	[__N __Y, to:_____]
4.2 Clarity of information need: __clear __medium __muddled __varies	[__N/K __N/A]	[__N __Y, to:_____]
4.3 Flow direction: __system to user __user to system __balanced __varies	[__N/K __N/A]	[__N __Y, to:_____]
4.4 Information volume: __low __medium __high __specific: _____	[__N/K __N/A]	[__N __Y, to:_____]
4.5 Complexity of information: __low __medium __high __varies	[__N/K __N/A]	[__N __Y, to:_____]
4.6 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]

5. INTERACTION

5.1 Locus of control: __push (system) __pull (user) __varies [__N/K __N/A] [__N __Y, to:_____]

5.2 Complexity of interaction: __low __medium __high __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

5.3 Predictability of interaction: __low __medium __high __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

5.4 Frequency: __rare __recurrent __frequent __varied __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

5.5 Regularity: __irregular__ regular period __varied __specific:_____ [__N/K __N/A] [__N __Y, to:_____]

5.6 Goal-orientation: __random __vague __average __goal oriented __other:_____ [__N/K __N/A] [__N __Y, to:_____]

5.7 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

6. ORIENTATION

6.1 Motivation: __high __middle __low __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

6.2 Likelihood of changing role: __low __medium __high __specific: _____

To what roles and when? _____ [__N/K __N/A] [__N __Y, to:_____]

6.3 Likelihood of abandoning system: __low __medium __high __specific: _____

On what conditions or why?_**on repeated perception of low performance & external factors**_____ [__N/K __N/A] [__N __Y, to:_____]

6.4 Purpose of use: __Professional __leisure/utility __leisure/entertainment __other: _____ [__N/K __N/A] [__N __Y, to:_____]

6.5 Optionality of use: __ Required use __optional use (conditions): _____ [__N/K __N/A] [__N __Y, to:_____]

6.6 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

7. RESTRICTIONS

7.1 Cost of Errors: __low __medium __high __specific: _**varies**_____ [__N/K __N/A] [__N __Y, to:_____]

7.2 Time restrictions: __none __low __medium __high __specific: _**varies**_____ [__N/K __N/A] [__N __Y, to:_____]

7.3 Access restrictions: __none __confidentiality/access rights __pay-per-view __pay-per-search __pay-per-time

__other:_____ [__N/K __N/A] [__N __Y, to:_____]

7.4 Device restrictions: __size __processing speed __available other tools or programs __input means __output means

__other _____ [__N/K __N/A] [__N __Y, to:_____]

7.5 Network restrictions: __low __medium __high __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

7.6 Additional restrictions and requirements related to organizational context (e.g., coverage requirements, etc.):

__**coverage needs to be complete within specific time period** _____ [__N/K __N/A] [__N __Y, to:_____]

7.7 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

8. PHYSICAL ENVIRONMENT

8.1 Mobility: __mobile __stationary __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

8.2 Geo-position: __one __many __specific:_____ [__N/K __N/A] [__N __Y, to:_____]
8.3 Distractions: __noise __interruptions __parallel tasks __other:_____ [__N/K __N/A] [__N __Y, to:_____]
8.4 Climate and lighting conditions: __lighting __temperature __humidity __other:_____ [__N/K __N/A] [__N __Y, to:_____]
8.5 Additional features /notes:_____ [__N/K __N/A] [__N __Y, to:_____]

9. SUCCESS CRITERIA

9.1 Efficiency: _4_ [__N/K __N/A] [__N __Y, to:_____]
9.2 Effectiveness: _1_ [__N/K __N/A] [__N __Y, to:__9__]
9.3 Satisfaction: _7_ [__N/K __N/A] [__N __Y, to:_____]
9.4 System reliability: _3_ [__N/K __N/A] [__N __Y, to:_____]
9.5 System intuitiveness: _5_ [__N/K __N/A] [__N __Y, to:_____]
9.6 System comprehensibility: _2_ [__N/K __N/A] [__N __Y, to:_____]
9.7 Actionability: _6_ [__N/K __N/A] [__N __Y, to:_____]
9.8 Additional criteria: ___ [__N/K __N/A] [__N __Y, to:_____]
9.9 Notes on success criteria:

_____ A searcher will usually stop when she has gathered enough information to clearly support a decision. How quickly she arrives to this is also one success criteria (Efficiency) _____
_____

Form 2 – Interaction and goals

1. USE CASE NAME AND SUPPORTED USER ROLES

1.1 Name:_____**claimValidity**_____

1.2 Supports (user roles):_____**Searcher** _____

2. USER GOALS	Not Know/ Applicable	Related to Evaluation
2.1 Type of information: __single fact/answer/notification __collection of facts/answers/notifications __single item (e.g., document) __collection of items __other:_____	[__N/K __N/A]	[__N __Y, to:_2.2]
2.2 Type of goal: __viewing __exporting __navigating __ordering/buying (transactional) __manipulating __surfing __other:_____	[__N/K __N/A]	[__N __Y, to:_____]

3. USE CASE RELATIONSHIPS – not the case yet, not defined

3.1 Specializes:_**priorArtSearch**_____	[__N/K __N/A]	[__N __Y, to:_____]
3.2 Extends:_____	[__N/K __N/A]	[__N __Y, to:_____]
3.3 Uses:_____	[__N/K __N/A]	[__N __Y, to:_____]
3.4 Resembles:_____	[__N/K __N/A]	[__N __Y, to:_____]

4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

claimValidity	EXTENDS:
USER INTENTION	SYSTEM RESPONSIBILITY
Request documents w.r.t. a set of given patent claims	
	Show ranked list of relevant results with text snippets, metadata information, links to full documents
Click on one element of the list	
	Display the full document with any metadata, attached images and text
Confirm document as relevant, assign a relevance degree	
	Save document and relevance to a list of user-selected documents
Go back to the original list of results	

	Display the first ranked list of results, the viewed one visibly identifiable
[continue selecting/clicking documents in the list until satisfied/found enough results] Save list of selected documents	
	Export the user-selected documents and save it at a user specified location Close (Use Case Ends)
EXTENSIONS: **priorArtSearch**	

Form 3 - System and interface feature checklist

	Not Known/ Applicable	Related to Evaluation
1. REQUEST FORMULATION		
1.1 Supported search strategies:		
__querying __browsing __monitoring __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.2 Query persistence: __one shot __permanent __evolving __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.3 Query modality: __text __image __video __audio __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.4 Query formulation: __specification __example __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.5 Query language:		
__simple keyword __basic operators __advanced __specific:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.6 Query target: __content __metadata/description __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.7 Query support:		
__spelling correction __query suggestion __translation __advanced query fields (support for advanced query language __QE __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.8 Browsing (content) categories:		
__people __country __subject __media __date __period __language __collection __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.9 Navigation support:		
__sitemap __FAQ __classifications __thesauri __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.10 Changing between querying and browsing:		
__supported __not supported __specific:_____	[__N/K __N/A]	[__N __Y, to:_____]
1.11 Additional query formulation features/notes:_____	[__N/K __N/A]	[__N __Y, to:_____]
2. RESULT PRESENTATION		
2.1 Presentation hierarchy: __one level __two level __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
2.2 Presentation granularity:		
__title __summary __metadata __full information item __set of items __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
2.3 Presentation organization:		
__single item __multiple items __list __ranked list __browsing interface __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
2.4 Result ordering (by):		

__score __date __diversity __author __random __other: __**subject, authority** _____ [__N/K __N/A] [__N __Y, to:__10_]

2.5 Assessment support:

__highlighting __scores __popularity __number of results __relations within a document

__relations between documents __other: _____ [__N/K __N/A] [__N __Y, to:_____]

2.6 Additional result presentation features/notes:_____ [__N/K __N/A] [__N __Y, to:_____]

3. RESULT USE

3.1 Manipulation:

__tagging __annotation __commenting __discussing __creating lists of documents __other:_____ [__N/K __N/A] [__N __Y, to:_10.1]

3.2 On site consumption/use: __viewing (on screen)__listening (within the system) __analysis and interpretation

__Other:_____ [__N/K __N/A] [__N __Y, to:_____]

3.3 Exporting search context [queries/number of results]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K __N/A] [__N __Y, to:_____]

3.4 Exporting results [single items/sets of documents]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K __N/A] [__N __Y, to:_____]

3.5 Sharing: __within the system __exporting __other:_____ [__N/K __N/A] [__N __Y, to:_____]

3.6 Ordering/paying: __internal __external __other: _____ [__N/K __N/A] [__N __Y, to:_____]

Evaluation – Form 4

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down on the "whiteboard" under each section.

1. TEST PERSON CHARACTERISTICS [__ Not applicable] **RELEVANT U.C FEATURES**

1.1 Are test persons representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Test persons are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task (e.g. motivation)

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

THE WHITEBOARD

2. TOPICS [__Not applicable]

2.1 Are the topics representative of the real search topics/information needs?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F1: 2.1, 2.4]

Topics representative of the search topics/real information needs in terms of, e.g.:

__domain of topics __clarity of information need (clear/muddled): __type of search goal

__information need durability.

__other:_____ [__N __Y, to:_F1: 4.2___]

__Real information needs/search topics not known/well understood

2.2 The topics are used for

__relevance assessments __automatic runs __tasks given to test persons

__other:_____ [__N __Y, to:_F1: 2.1___]

THE WHITEBOARD

3. REQUESTS [__Not applicable]

3.1 Are the requests representative of the real requests?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F1: 5.6___]

The requests are representative of the real requests in terms of, e.g.:

__type of request __request modality __query length __query quality __query structure

__query formulation (specification/example) __query durability

__other:_____ [__N __Y, to:_ F1: 5.6___]

__Real requests not known

THE WHITEBOARD

4. DATA [__Not applicable]

4.1 Is the data used in the evaluation activity representative of the real data?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F1: 3.3___]

The test data used is representative of the real data in terms of e.g.:

__intellectual content __modality __size __dynamics __curation __provenance (created by one/created by many)

__structure __granularity

__other:_____ [__N __Y, to:_F1: 3_____]

__Real data not known

THE WHITEBOARD

5. GROUND TRUTH CREATION [__Not applicable]

5.1 Ground truth captures:

__relevance of documents to topics __which of two documents is more relevant to a topic

__which of two ranked lists is better __other:_____ [__N __Y, to:_____]

5.2 Ground truth is obtained:

__ manually __semi-automatically (e.g. pooling) __fully automatically __other:_____ [__N __Y, to: _____]

5.3 Are the relevance criteria representative of the real users' relevance criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Relevance criteria are representative of the real users' relevance criteria in terms of e.g.:

__strictness of criteria __type of criteria (e.g. topicality or novelty) __grades of relevance

__other:_____ [__N __Y, to:_____]

__Real users' relevance criteria not known/well understood

5.4 Are assessors representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Assessors are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

5.5 Are the results shown to assessors representative of the real results shown to end users?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

n this Use Case we use the existing search reports done by patent experts in their professional activity.

6. RESULT PRESENTATION [__Not applicable]

6.1 Is the result presentation in experiment representative of target system(s)?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F3: 2 ____]

The result presentation is representative of target system(s) in terms of e.g.:

__presentation hierarchy __granularity __other:__**ranked list** _____ [__N __Y, to: _ F3: 2 ____]

__Target system result presentation not known

THE WHITEBOARD

7. INTERACTION [__Not applicable]

7.1 Interaction in the experiment is:

__real user interaction __interaction model __other:_____ [__N __Y, to: _____]

7.2 Is the interaction in the experiment representative of real end user-system interaction?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The interaction is representative of real end user-system interaction in terms of e.g.:

__search strategies __result assessment __goal orientation __learning

__session length/complexity __query reformulation (strategies, cost, …)

__other:_____ [__N __Y, to: _____]

__Real interaction patterns not known/well understood

THE WHITEBOARD

8. RESULT USE [__Not applicable]

8.1 Result use is included in evaluation with:

__real users, real use __ real users, controlled use __simulated __no result use_____ [__N __Y, to: _____]

8.2 Is the result use in the experiment representative of the real result use patterns?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result use is representative of the real result use in terms of, e.g.:

__type of use/search goals __effect on success criteria __effect on information needs

__other:_____ [__N __Y, to: _____]

__the result use of end users is not known/well understood.

THE WHITEBOARD

This Use Case doesn't evaluate interactive systems

9. EVALUATION CRITERIA [__Not applicable]

9.1 Are the success criteria in the experiment representative of end users' success criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F1: 9 ____]

The success criteria are representative of end users success criteria in terms of, e.g.:

__volume of relevant results __time spent (if the goal is to spend time) __user satisfaction

__meeting user expectations __task completion __objectivity/subjectivity of criteria

__other:_____ [__N __Y, to: _F1: 9.2___]

__End users' success criteria not known/well-understood

__Evaluation is not based on user criteria, but:_____

9.2 Are the failure criteria in the experiment representative of end users' failure criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _F1: 9.2 __]

The failure criteria are representative of end users' failure criteria in terms of, e.g.:

__time __effort __poor result quality __frustration __"out of queries" __other: _____ [__N __Y, to: _F1: 9.2 __]

__End users' failure criteria not known/well understood

10. METRICS [__Not applicable]

10.1 Do the metrics measure what matters most to the end users?

__yes __reasonably __no, not very __ that's not the goal __not known

__other:_____ [__N __Y, to: _F3: 3.1 __]

Metrics measure what matters most to the end users in terms of, e.g.:

__task completion __cost of errors __efficiency __time spent __effort __domain restrictions

__other:_____ [__N __Y, to: _F3: 3.1 __]

10.2 Are the metrics used predictive of real world performance?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The metrics are predictive of real world performance, in terms of, e.g.:

__relative performance (between systems) __absolute performance __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

Appendix H. Search for lecture material checklists

Background Feature Checklist – Form 1 (Search for Cultural Heritage)

1. USER ROLE

1.1 Role Name____**Searcher** _____

1.2 Related Roles___**Browser, Flaneur**_____

	Not known/ not applicable	Related to evaluation
2. INCUMBENTS		
2.1 Domain knowledge: __ none __limited **x** moderate __high __varies	[__N/K __N/A]	[__N **x** Y, to: ____**2**]
2.2 General search or system proficiency: __ novice _ **x** _moderate __expert __varies	[__N/K __N/A]	[__N **x** Y, to: ____ **2**]
2.3 System knowledge: __ none **x** limited __moderate __high __varies	[__N/K __N/A]	[__N **x** Y, to: ____ **2**]
2.4 Language skills: __ none __limited __moderate **x** high __varies	[__N/K __N/A]	[__N **x** Y, to: ____ **2**]
2.5 Other relevant user features (e.g., age, training, education, disabilities, etc.): __**deep knowledge of subject area**____	[__N/K __N/A]	[__N **x** Y, to:____ **2**]
2.6 Additional features/notes: _____	[**x** N/K __N/A]	[**x** N __Y, to:_____]
3. REPOSITORY		
3.1 Media: **x** text **x** image **x** video **x** audio __ graphs __3D objects __ varies __other: _____	[__N/K __N/A]	[__N **x** Y, to:____**4**]
3.2 Granularity (of what is an information item): __low __ medium **x** high __varies __specific: _____	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
3.3 Genre: __news __factual __entertainment **x** scientific __commercial __personal commentary __technical text __varies __other: __**cultural heritage**_____	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
3.4 Language: __monolingual __bilingual **x** multilingual __ other: _____	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
3.5 Technical Quality: **x** low __moderate __high __varies	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
3.6 Source Dynamics: __static collection **x** dynamic collection __stream __other: _____	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
3.7 Indexing Timeliness: __immediate __every hour __daily **x** weekly __monthly __varies __other: _____	[__N/K __N/A]	[**x** N __Y, to:_____]
3.8 Additional features/notes: _____	[**x** N/K __N/A]	[**x** N __Y, to:_____]
4. INFORMATION		
4.1 Origin of user input: __aural __visual **x** mental __touch __varies __specific:_____	[__N/K __N/A]	[**x** N __Y, to:_____]
4.2 Clarity of information need: **x** clear __medium __muddled __varies	[__N/K __N/A]	[__N **x** Y, to:____**3**]
4.3 Flow direction: **x** system to user __user to system __balanced __varies	[__N/K __N/A]	[__N **x** Y, to:__**3,4,7**]
4.4 Information volume: __low **x** medium __high __specific: _____	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
4.5 Complexity of information: **x** low __medium __high __varies	[__N/K __N/A]	[__N **x** Y, to:____ **4**]
4.6 Additional features/notes: _____	[**x** N/K __N/A]	[**x** N __Y, to:_____]

5. INTERACTION

5.1 Locus of control: __push (system) x pull (user) __varies [__N/K __N/A] [__N x Y, to:____ 7]

5.2 Complexity of interaction: x low __medium __high __varies __specific: _____ [__N/K __N/A] [__N x Y, to:____ 7]

5.3 Predictability of interaction: __low __medium x high __specific: _____ [__N/K __N/A] [__N x Y, to:____ 7]

5.4 Frequency: __rare x recurrent __frequent __varied __specific: _____ [__N/K __N/A] [x N __Y, to:_____]

5.5 Regularity: x irregular__ regular period __varied __specific:_____ [__N/K __N/A] [x N __Y, to:_____]

5.6 Goal-orientation: __random __vague __average x goal oriented __other:_____ [__N/K __N/A] [__N x Y, to:____ 7]

5.7 Additional features/notes: _____ [x N/K __N/A] [x N __Y, to:_____]

6. ORIENTATION

6.1 Motivation: x high __middle __low __varies __specific: _____ [__N/K __N/A] [x N __Y, to:_____]

6.2 Likelihood of changing role: __low x medium __high __specific: _____

To what roles and when? __**from searcher to browser – only after initial search**_____ [__N/K __N/A] [x N __Y, to:__ 7, 8]

6.3 Likelihood of abandoning system: __low __medium x high __specific: _____

On what conditions or why?___**when no relevant objects are found**_____ [__N/K __N/A] [__N x Y, to:____ 8]

6.4 Purpose of use: x Professional __leisure/utility __leisure/entertainment __other: _____ [__N/K __N/A] [__N x Y, to:____ 7]

6.5 Optionality of use: __ Required use x optional use (conditions): _____ [__N/K __N/A] [x N __Y, to:_____]

6.6 Additional features/notes: _____ [x N/K __N/A] [x N __Y, to:_____]

7. RESTRICTIONS

7.1 Cost of Errors: x low __medium __high __specific: _____ [__N/K __N/A] [x N __Y, to:_____]

7.2 Time restrictions: __none x low __medium __high __specific: _____ [__N/K __N/A] [x N __Y, to:_____]

7.3 Access restrictions: x none __confidentiality/access rights __pay-per-view __pay-per-search __pay-per-time __other:_____ [__N/K __N/A] [__N x Y, to:____ 4]

7.4 Device restrictions: __size __processing speed __available other tools or programs __input means _output means __other _____ [__N/K x N/A] [x N __Y, to:_____]

7.5 Network restrictions: x low __medium __high __varies __specific: _____ [__N/K __N/A] [__N x Y, to:____ 7]

7.6 Additional restrictions and requirements related to organizational context (e.g., coverage requirements, etc.):

_____ [x N/K __N/A] [x N __Y, to:_____]

7.7 Additional features/notes: _____ [x N/K __N/A] [x N __Y, to:_____]

8. PHYSICAL ENVIRONMENT

8.1 Mobility: __mobile __stationary x varies __specific: _____ [__N/K __N/A] [x N __Y, to:_____]

8.2 Geo-position: x one __many __specific:_____ [__N/K __N/A] [x N __Y, to:_____]
8.3 Distractions: __noise x interruptions __parallel tasks __other:_____ [__N/K __N/A] [x N __Y, to:_____]
8.4 Climate and lighting conditions: __lighting __temperature __humidity __other:_____ [__N/K x N/A] [x N __Y, to:_____]
8.5 Additional features /notes:_____ [x N/K __N/A] [x N __Y, to:_____]

9. SUCCESS CRITERIA

9.1 Efficiency: ___ [__N/K x N/A] [x N __Y, to:_____]
9.2 Effectiveness: _ x _ [__N/K __N/A] [__N x Y, to:__9,10]
9.3 Satisfaction: _ x _ [__N/K __N/A] [x N __Y, to:_____]
9.4 System reliability: ___ [__N/K x N/A] [x N __Y, to:_____]
9.5 System intuitiveness: _ x _ [__N/K __N/A] [x N __Y, to:_____]
9.6 System comprehensibility: _ x _ [__N/K __N/A] [x N __Y, to:_____]
9.7 Actionability: ___ [__N/K x N/A] [x N __Y, to:_____]
9.8 Additional criteria: ___ [__N/K x N/A] [x N __Y, to:_____]
9.9 Notes on success criteria:
____Efficiency and reliability are only relevant if the system breaks down completely, otherwise there is flexibility wrt success experience for user
_____
_____
_____

Form 2 – Interaction and goals – Search for cultural heritage

1. USE CASE NAME AND SUPPORTED USER ROLES

1.1 Name:___**Search for cultural heritage material**_____

1.2 Supports (user roles):_____**Searcher, browser, flaneur**_____

	Not Know/ Applicable	Related to Evaluation
2. USER GOALS		
2.1 Type of information: __single fact/answer/notification __collection of facts/answers/notifications __single item (e.g., document) **x** collection of items __other:_____ _____	[__N/K __N/A]	[__N **x** Y, to:____ **8**]
2.2 Type of goal: x viewing **x** exporting __navigating __ordering/buying (transactional) __manipulating __surfing __other:_____	[__N/K __N/A]	[__N **x** Y, to:____**8**]

3. USE CASE RELATIONSHIPS

3.1 Specializes:_____	[__N/K **x** N/A]	[__N __Y, to:_____]
3.2 Extends:_____	[__N/K **x** N/A]	[__N __Y, to:_____]
3.3 Uses:_____ **simple ad-hoc search, simple browse**_____	[__N/K __N/A]	[__N **x** Y, to:____ **7**]
3.4 Resembles:_____**simple ad-hoc search, simple browse**_____	[__N/K __N/A]	[__N **x** Y, to:____ **7**]

4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

Search for CH Material	EXTENDS: Ad-hoc search
USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction	
Type query	Present result list
Choose facet	Reduce result list according to facet characteristic
Browse result pages	Enable paging of result list
View individual object	Open individual object page
Click related object	Open individual object page
Click external link	Enable link tracing
	Close (Use Case Ends)
EXTENSIONS	

Form 3 - System and interface feature checklist – Search for cultural heritage

	Not Known/ Applicable	Related to Evaluation
1. REQUEST FORMULATION		
1.1 Supported search strategies:		
x querying x browsing __monitoring __other:_____	[__N/K __N/A]	[__N x Y, to:____**3,7**]
1.2 Query persistence: x one shot __permanent __evolving __other:_____	[__N/K __N/A]	[__N x Y, to:_____ **3**]
1.3 Query modality: _ x text __image __video __audio __other:_____	[__N/K __N/A]	[__N x Y, to:____ **3**]
1.4 Query formulation: x specification __example __other:_____	[__N/K __N/A]	[__N x Y, to:____ **3**]
1.5 Query language:		
x simple keyword __basic operators __advanced __specific:_____	[__N/K __N/A]	[__N x Y, to:____ **3**]
1.6 Query target: __content x metadata/description __other:_____	[__N/K __N/A]	[__N x Y, to:____ **3**]
1.7 Query support:		
__spelling correction x query suggestion __translation __advanced query fields (support for advanced query language __QE __other:_____	[__N/K __N/A]	[__N x Y, to:___**3,7**]
1.8 Browsing (content) categories:		
__people x country __subject x media x date __period x language x collection __other:_____	[__N/K __N/A]	[x N __Y, to:_____]
1.9 Navigation support:		
x _sitemap x _FAQ __classifications __thesauri __other:_____	[__N/K __N/A]	[x N __Y, to:_____]
1.10 Changing between querying and browsing:		
x supported __not supported __specific:_____	[__N/K __N/A]	[x N __Y, to:_____]
1.11 Additional query formulation features/notes:_____	[x N/K __N/A]	[x N __Y, to:_____]
2. RESULT PRESENTATION		
2.1 Presentation hierarchy: __one level x two level __other:_____	[__N/K __N/A]	[__N x Y, to:____**6**]
2.2 Presentation granularity:		
x title __summary __metadata __full information item __set of items __other:___**thumbnails**_____	[__N/K __N/A]	[__N x Y, to:____ **6**]
2.3 Presentation organization:		
__single item __multiple items __list x ranked list __browsing interface __other:_____	[__N/K __N/A]	[__N x Y, to:____ **6**]
2.4 Result ordering (by):		

x score __date __diversity __author __random __other:_____ [__N/K __N/A] [__N x Y, to:____ 6]

2.5 Assessment support:

__highlighting __scores __popularity __number of results __relations within a document

__relations between documents __other: _____ [__N/K x N/A] [x N __Y, to:_____]

2.6 Additional result presentation features/notes:___matrix of 4x3 thumbnails + title_____ [__N/K __N/A] [x N __Y, to:_____]

3. RESULT USE

3.1 Manipulation:

x tagging x annotation __commenting __discussing __creating lists of documents __other:_____ [__N/K __N/A] [x N __Y, to:_____]

3.2 On site consumption/use: x viewing (on screen)__listening (within the system) __analysis and interpretation

__Other:_____ [__N/K __N/A] [__N x Y, to:____ 6]

3.3 Exporting search context [queries/number of results]:

x saving x printing x publishing (social media, etc.) __other:_____ [__N/K __N/A] [x N __Y, to:_____]

3.4 Exporting results [single items/sets of documents]:

x saving x printing x publishing (social media, etc.) __other:_____ [__N/K __N/A] [x N __Y, to:_____]

3.5 Sharing: __within the system x exporting __other:_____ [__N/K __N/A] [x N __Y, to:_____]

3.6 Ordering/paying: __internal __external __other: _____ [__N/K x N/A] [x N __Y, to:_____]

Evaluation – Form 4 – Search for cultural heritage

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down on the "whiteboard" under each section.

1. TEST PERSON CHARACTERISTICS [x Not applicable] **RELEVANT U.C FEATURES**

1.1 Are test persons representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Test persons are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task (e.g. motivation)

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

2. TOPICS [__Not applicable]

2.1 Are the topics representative of the real search topics/information needs?

__yes **x** reasonably __no, not very __not known __other:_____ [__N **x** Y, to: _____**2.2**]

Topics representative of the search topics/real information needs in terms of, e.g.:

x domain of topics **x** clarity of information need (clear/muddled): **x** type of search goal

x information need durability.

__other:_____ [__N **x** Y, to:_____ **2.2**]

x Real information needs/search topics not known/well understood

2.2 The topics are used for

x relevance assessments __automatic runs __tasks given to test persons

__other:_____ [__N **x** Y, to:_____ **2.2**]

3. REQUESTS [__Not applicable]

3.1 Are the requests representative of the real requests?

x yes __reasonably __no, not very __not known __other:_____ [__N x Y, to: _____**2.2**]

The requests are representative of the real requests in terms of, e.g.:

x type of request x request modality x query length x query quality x query structure

x query formulation (specification/example) x query durability

__other:_____ [__N x Y, to:_____ **2.2**]

__Real requests not known

4. DATA [__Not applicable]

4.1 Is the data used in the evaluation activity representative of the real data?

x yes __reasonably __no, not very __not known __other:_____ [__N x Y, to: _____**1.3**]

The test data used is representative of the real data in terms of e.g.:

x intellectual content x modality x size __dynamics x curation x provenance (created by one/created by many)

x structure __granularity

__other:_____ [__N __Y, to:_____ **1.3**]

__Real data not known [__N __Y, to:_____ **1.3**]

5. GROUND TRUTH CREATION [__Not applicable]

5.1 Ground truth captures:

x relevance of documents to topics __which of two documents is more relevant to a topic

__which of two ranked lists is better __other:_____ [__N x Y, to:_____**1.6, 2.2**]

5.2 Ground truth is obtained:

x manually __semi-automatically (e.g. pooling) __fully automatically __other:_____ [__N x Y, to: _____**1.6, 2.2**]

5.3 Are the relevance criteria representative of the real users' relevance criteria?

__yes x reasonably __no, not very __not known __other:_____ [__N x Y, to: _____**1.6, 2.2**]

Relevance criteria are representative of the real users' relevance criteria in terms of e.g.:

x strictness of criteria x type of criteria (e.g. topicality or novelty) __grades of relevance

__other:_____ [__N x Y, to:____ 1.6, 2.2]

__Real users' relevance criteria not known/well understood

5.4 Are assessors representative of the end user population?

__yes x reasonably __no, not very __not known __other:_____ [__N x Y, to: _____1.2]

Assessors are representative of the end user population in terms of, e.g.:

__demographics x search skills x domain knowledge x language skills __relation to search task

__other:_____ [__N __Y, to:_____ 1.2]

__ End user population not known/well understood

5.5 Are the results shown to assessors representative of the real results shown to end users?

__yes x reasonably __no, not very __not known __other:_____ [__N x Y, to: _____3.2]

6. RESULT PRESENTATION [__Not applicable]

6.1 Is the result presentation in experiment representative of target system(s)?

__yes x reasonably __no, not very __not known __other:_____ [__N __Y, to: _____3.2]

The result presentation is representative of target system(s) in terms of e.g.:

__presentation hierarchy x granularity __other:_____ [__N __Y, to: _____3.2]

__Target system result presentation not known

7. INTERACTION [x Not applicable]

7.1 Interaction in the experiment is:

__real user interaction __interaction model __other:_____ [__N __Y, to: _____]

7.2 Is the interaction in the experiment representative of real end user-system interaction?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The interaction is representative of real end user-system interaction in terms of e.g.:

__search strategies __result assessment __goal orientation __learning

__session length/complexity __query reformulation (strategies, cost, …)

__other:_____ [__N __Y, to: _____]

__Real interaction patterns not known/well understood

8. RESULT USE [x Not applicable]

8.1 Result use is included in evaluation with:

__real users, real use __ real users, controlled use __simulated __no result use_____ [__N __Y, to: _____]

8.2 Is the result use in the experiment representative of the real result use patterns?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result use is representative of the real result use in terms of, e.g.:

__type of use/search goals __effect on success criteria __effect on information needs

__other:_____ [__N __Y, to: _____]

__the result use of end users is not known/well understood.

9. EVALUATION CRITERIA [__Not applicable]

9.1 Are the success criteria in the experiment representative of end users' success criteria?

__yes x reasonably __no, not very __not known __other:_____ [__N x Y, to: _____1.9]

The success criteria are representative of end users success criteria in terms of, e.g.:

x volume of relevant results __time spent (if the goal is to spend time) __user satisfaction

x meeting user expectations __task completion __objectivity/subjectivity of criteria

__other:_____ [__N x Y, to: _____1.9]

__End users' success criteria not known/well-understood

__Evaluation is not based on user criteria, but:_____

9.2 Are the failure criteria in the experiment representative of end users' failure criteria?

__yes __reasonably __no, not very x not known __other:_____ [__N x Y, to: _____1.6]

The failure criteria are representative of end users' failure criteria in terms of, e.g.:

__time __effort __poor result quality __frustration __"out of queries" x other: __not considered_____ [__N x Y, to: _____1.6]

__End users' failure criteria not known/well understood

10. METRICS [__Not applicable]

10.1 Do the metrics measure what matters most to the end users?

__yes **x** reasonably __no, not very __ that's not the goal __not known

__other:_____ [__N **x** Y, to: _____**1.9**]

Metrics measure what matters most to the end users in terms of, e.g.:

x task completion __cost of errors __efficiency __time spent __effort __domain restrictions

__other:_____ [__N **x** Y, to: _____**1.9**]

10.2 Are the metrics used predictive of real world performance?

__yes **x** reasonably __no, not very __not known __other:_____ [__N **x** Y, to: _____**1.9**]

The metrics are predictive of real world performance, in terms of, e.g.:

x relative performance (between systems) __absolute performance __other:_____ [__N **x** Y, to: _____**1.9**]

Appendix I. Reputation management checklists

multilingual

Background Feature Checklist – Form 1

Problem Statement: Evaluating reputation man. sys. that clusters tweets referring to a brand, and then rank these clusters ~~as~~ by priority, where priority means the priority that a company executive should act on these clusters

Role Name reputation – analyst.

Related Roles company executives (CEOs)

value, ICES, ... information analyst
Google midway
reputation man. "middle ma" (ddlC) (expert analyst.

	Not known/ not applicable	Related to evaluation
INCUMBENTS (Actual users in role)		
Domain Knowledge: __ none __limited __moderate X high __varies	[__N/K __N/A]	[__N __Y, to:____]
General Search Proficiency: __ none __limited X moderate __high __varies	[__N/K X N/A]	[__N __Y, to:____]
System Knowledge: __ none __limited X moderate __high __varies	[__N/K __N/A]	[__N __Y, to:____]
Language Skills: __ none __limited __moderate X high __varies	[__N/K __N/A]	[__N __Y, to:____]
Information Need Definition: X clear __medium __muddled __varies	[__N/K __N/A]	[__N __Y, to:____]
Other Relevant User Features (e.g., age, training, education, disabilities, etc.): professionals (paid to do task)	[__N/K __N/A]	[__N __Y, to:____]
REPOSITORY		
Media: X text __image X video X audio __ graphs __3D objects X varies __other:____	[__N/K __N/A]	[__N __Y, to:____]
Genre: X news X factual __entertainment X scientific __commercial X personal commentary __technical text X varies __other: fora, blogs	[__N/K __N/A]	[__N __Y, to:____]
Language: __monolingual __bilingual X multilingual __ other:____	[__N/K __N/A]	[__N __Y, to:____]
Technical Quality: __low __moderate __high X varies	[__N/K __N/A]	[__N __Y, to:____]
Source Dynamics: __collection X stream __other:____	[__N/K __N/A]	[__N __Y, to:____]
Indexing Timeliness: X immediate __every hour __daily __weekly __monthly __varied __other: minutely!	[__N/K __N/A]	[__N __Y, to:____]
INFORMATION		
Volume Exchanged (between user and system): __low X medium __high __specific:____	[__N/K __N/A]	[__N __Y, to:____]
Relevant volume (potentially of interest in repository): __low __medium __high __specific: varies	[__N/K __N/A]	[__N __Y, to:____]
Granularity (of what is an information item): __data element (low) __ (medium) X (high) X specific: utterances	[__N/K __N/A]	[__N __Y, to:____]
Complexity of information: __low __medium X high __varies	[__N/K __N/A]	[__N __Y, to:____]

↳ depends: - utterances in isolation are simple
- aggregated understanding is complex.

not really applicable

brand name

schedule, do we measure this?

INTERACTION

Origin of user input: X external __~~thoughts~~ __varies __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Flow (predominant direction): X system to user X user to system __balanced X varies [__N/K __N/A] [__N __Y, to:____]

Locus of control: X push (system) X pull (user) X varies ? I would like alarm bells going off [__N/K __N/A] [__N __Y, to:____]

Complexity of interaction: X low __medium __high __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Predictability of interaction: __low __medium X high __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Frequency: __rare __recurrent __frequent __varied __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Regularity: __sporadic __regular __varied __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Continuity: __intermittent __continuous __varied __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Goal-orientation: __random __vague __average __goal oriented __other: _____ [__N/K __N/A] [__N __Y, to:____]

ORIENTATION

Valence (towards system or system use): __positive __negative __ambivalent X varies

__other: not sure if experts would trust systems yet (good enough?) [__N/K __N/A] [__N __Y, to:____]

Motivation (towards role/interaction): X high __middle __low __varies

__specific: to professional paid [__N/K __N/A] [__N __Y, to:____]

Emotion: __frustrated X anxious X neutral __happy X varies __other: work attitude [__N/K __N/A] [__N __Y, to:____]

Likelihood of changing role: X low __medium __high __specific %: _____

To what roles and when? _____ [__N/K __N/A] [__N __Y, to:____]

Likelihood of abandoning system: __low X medium __high __specific %: _____

On what conditions or why? quality: are items missing? [__N/K __N/A] [__N __Y, to:____]

Purpose of use: X Professional __leisure/utility __leisure/entertainment __other: _____ [__N/K __N/A] [__N __Y, to:____]

__ Required use X optional use (conditions): other systems are available [__N/K __N/A] [__N __Y, to:____]

RESTRICTIONS

Cost of Errors: __low __medium X high __specific: missing a threat: high cost [__N/K __N/A] [__N __Y, to:____]

Urgency: __low __medium X high __specific: _____ [__N/K __N/A] [__N __Y, to:____]

Access restrictions (confidentiality etc.): __none __task related __user related

__other: well, harvesting content might be a legal issue? [__N/K __N/A] [__N __Y, to:____]

Cost (time/money): X no cost __time restrictions __pay-per-view __pay-per-search

__other: _____ [__N/K __N/A] [__N __Y, to:____]

a repository feature?

cost for whom? → user, → content provider, → ...

Device: X table-top __laptop __smart phone __game console __e-book device __varies
__other:_____ [__N/K __N/A] [__N __Y, to:____]
Input means: X keyboard __keypad __touch pad __mouse __microphone __camera __varies
__other:_____ [__N/K __N/A] [__N __Y, to:____]
Output means: __small screen X suitable screen __speakers __earplugs __varies
__other:_____ [__N/K __N/A] [__N __Y, to:____]
Other restrictions and requirements related to organizational context (e.g., coverage requirements, etc.):
_____ [__N/K __N/A] [__N __Y, to:____]
Network (latency/cost): __low X medium __high __varies __specific:_____ [__N/K __N/A] [__N __Y, to:____]

PHYSICAL ENVIRONMENT

Mobility: ? mobile X stationary __varies __specific: given time constraints, perhaps immobile should be supported? [__N/K __N/A] [__N __Y, to:____]
Geo-position: X one X many __specific: ? [__N/K __N/A] [__N __Y, to:____]
Other relevant features (e.g., lighting, noisiness etc.):
X [__N/K __N/A] [__N __Y, to:____]

X

FUNCTIONAL SUPPORT

Needed by users in this role: X X X X X
_____ [__N/K __N/A] [__N __Y, to:____]

SUCCESS CRITERIA

X system efficiency X system reliability __system intuitiveness X accuracy __coverage X flexibility of operation
X system comprehensibility (transparency) __clarity of presentation __user experience __satisfaction — conversion rate — revenue — market share
__other: high recall, but especially of high priority clusters/tweets. [__N/K __N/A] [__N __Y, to:____]

→ yes: clients want to know the why. (Magnus)

perhaps an additional metric! could be just measuring whether alarm bells go off when they should?

Use Case Model – Form 2

NAME: cluster-rank

SUPPORTS (user roles): rep-manager

USE CASE RELATIONSHIPS

	Not Know/ Applicable	Related to Evaluation
__Specializes:	[__N/K __N/A]	[__N __Y, to:____]
__Extends:	[__N/K __N/A]	[__N __Y, to:____]
__Uses:	[__N/K __N/A]	[__N __Y, to:____]
__Resembles:	[__N/K __N/A]	[__N __Y, to:____]

? this is the heart of it

GOAL

Type of information:

__single fact/answer/notification X collection of facts/answers/notifications

__single item (e.g., document) X collection of items __other: ranking clusters by priority [__N/K __N/A] [__N __Y, to:____]

Type of goal: X viewing __exporting __navigating __ordering/buying (transactional)

__other: deciding if action should be taken [__N/K __N/A] [__N __Y, to:____]

INTERACTION

usecaseName

USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction	
enter brand name	
	cluster tweets, rank clusters
decide if action should be taken.	
	Close (Use Case Ends)

EXTENSIONS

? out of scope of eval task.

TK: two users?
a junior searcher searches first
a senior searcher searches next

System and Interface Features – Form 3

	Not Known/ Applicable	Related to Evaluation
REQUEST FORMULATION		
Supported search strategies: X querying __browsing __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Query modality: X text __image __video __audio __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Query formulation: X specification __example __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Query language:		
X simple keyword __basic operators __advanced __specific:_____	[__N/K __N/A]	[__N __Y, to:_____]
Query target: X content __metadata/description __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Query support:		
__spelling correction __query suggestion __translation __advanced query fields (support for advanced query language __QE __other: none	[__N/K __N/A]	[__N __Y, to:_____]
Browsing (content) categories:		
__people _country __subject __media __date __period __language __collection __other: none	[__N/K __N/A]	[__N __Y, to:_____]
Navigation support:		
__sitemap __FAQ __classifications __thesauri __other: none	[__N/K __N/A]	[__N __Y, to:_____]
Changing between querying and browsing:		
__supported X not supported __specific:_____	[__N/K __N/A]	[__N __Y, to:_____]
RESULT PRESENTATION ✓ ? correct? → clusters, ranked		
Presentation hierarchy: __one level X two level __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Presentation granularity: → cluster labeling ?!		
__title __summary __metadata __full document X document set __other:_____	[__N/K __N/A]	[__N __Y, to:_____]
Presentation unit:		
__single item __multiple items __list __ranked list __browsing interface X other: list of clusters	[__N/K __N/A]	[__N __Y, to:_____]
Result organization (by): __score __date __diversity __author X other: priority	[__N/K __N/A]	[__N __Y, to:_____]
Assessment support:		
__highlighting __scores __popularity __number of results __relations within a document		
__relations between documents X other: none	[__N/K __N/A]	[__N __Y, to:_____]
Other result presentation aspects: _____	[__N/K __N/A]	[__N __Y, to:_____]

would be nice though: give the system some info and re-run.

RESULT USE

Manipulation:

__tagging __annotation __commenting __discussing __creating lists of documents X other: none [__N/K __N/A] [__N __Y, to:____]

~~**Viewing (on screen):** __Alt1 __Alt2 __Other:~~ [__N/K __N/A] [__N __Y, to:____]

Exporting search context [queries/number of results]:

__saving __printing __publishing (social media, etc.) X other: none [__N/K __N/A] [__N __Y, to:____]

Exporting results [single items/sets of documents]:

__saving __printing __publishing (social media, etc.) __other: [__N/K __N/A] [__N __Y, to:____]

Sharing: __within the system __exporting X other: none [__N/K __N/A] [__N __Y, to:____]

Ordering/paying: __Alt1 __Alt2 __Other: [__N/K __N/A] [__N __Y, to:____]

Stefan will provide us the Black box feedb. by next week.

Evaluation – Form 4

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down on the "whiteboard" under each section.

Related to RELEVANT U.C FEATURES

1. TEST PERSON CHARACTERISTICS [X Not applicable]

1.1 Are test persons representative of the end user population?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Test persons representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task (e.g. motivation)

__other:_____ [__N __Y, to:_____]

__ End user population not known/well understood

THE WHITEBOARD

/ = priority of topics in treats, to read or then

2. TOPICS [__Not applicable]

2.1 Are the topics representative of the real search topics/information needs?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Topics representative of the search topics/real information needs in terms of, e.g.:

X domain of topics X level of information need "definition" (clear/muddled): X type of search goal

__other:_____ [__N __Y, to:_____]

__Real information needs/search topics not known/well understood

2.2 The topics are used for ?

X relevance assessments __automatic runs __tasks given to test persons

__other:_____ [__N __Y, to:_____]

uneasy fit here

THE WHITEBOARD

3. REQUESTS [__Not applicable]

3.1 Are the requests representative of the real requests?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The requests are representative of the real requests in terms of, e.g.: (currently a bit)

X type of request X request modality __query length __query quality __query structure

__query formulation (specification/example)

__other: brand names [__N __Y, to:_____]

__Real requests not known

THE WHITEBOARD

4. DATA [__Not applicable]

4.1 Is the data used in the evaluation activity representative of the real data?

X yes X reasonably __no, not very __not known X other: its a subset. (Real data: also fora, other "neutral ground".) [__N __Y, to: _____]

The test data used is representative of the real data in terms of e.g.:

__contents __modality __size X dynamics __curation X provenance (created by one/created by many) [__N __Y, to:_____]

X structure __other:_____ [__N __Y, to:_____]

__Real data not known

THE WHITEBOARD

5. GROUND TRUTH CREATION [__Not applicable]

5.1 Relevance assessments capture:

X relevance of documents to queries __which of two documents is more relevant to a query

__which of two ranked lists is better X other: clustering of tweets, priority of cluster, in terms of many aspects. [__N __Y, to:_____]

5.2 Relevance assessments are obtained:

X manually __semi-automatically (e.g. pooling) __fully automatically __other:_____ [__N __Y, to: _____]

5.3 Are the relevance criteria representative of the real users' relevance criteria?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Relevance criteria are representative of the real users' relevance criteria in terms of e.g.:

X strictness of criteria X type of criteria (e.g. topicality or novelty) __grades of relevance

__other: actionable [__N __Y, to:_____]

__Real users' relevance criteria not known/well understood

5.4 Are assessors representative of the end user population?

X yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Assessors are representative of the end user population in terms of, e.g.:

__demographics __search skills X domain knowledge __language skills X relation to search task

__other: they are reputation managers who do this job for real. [__N __Y, to:_____]

__ End user population not known/well understood

THE WHITEBOARD

→ the system that assessors use

6. RESULT PRESENTATION [__Not applicable]

6.1 Is the result presentation in experiment representative of target system(s)?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result presentation is representative of target system(s) in terms of e.g.:

__presentation hierarchy __granularity __presentation unit x other: it's the same, but presentation issues are ignored [__N __Y, to: _____]

__Target system result presentation not known

THE WHITEBOARD

7. INTERACTION [X Not applicable]

7.1 Interaction in the experiment is:

__real user interaction __interaction model __other:_____ [__N __Y, to: _____]

7.2 Is the interaction in the experiment representative of real end user-system interaction?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The interaction is representative of real end user-system interaction in terms of e.g.:

__search strategies __result assessment __directness of interaction __learning

__session length/complexity __query reformulation (strategies, cost, ...)

__other:_____ [__N __Y, to: _____]

__Real interaction patterns not known/well understood

THE WHITEBOARD

8. SUCCESS AND FAILURE CRITERIA [X Not applicable]

8.1 Are the success criteria in the experiment representative of end users success criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The success criteria are representative of end users success criteria in terms of, e.g.:

__volume of relevant results __time spent (if the goal is to spend time) __user satisfaction

__meeting user expectations __task completion __objectivity/subjectivity of criteria – relevance

__other:_____ [__N __Y, to: _____]

__End users' success criteria not known/well-understood

8.2 Are the failure criteria in the experiment representative of end users success criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The failure criteria are representative of end users success criteria in terms of, e.g.:

__time __effort __poor result quality __frustration __"out of queries" __other: _____ [__N __Y, to: _____]

__End users' failure criteria not known/well understood

THE WHITEBOARD

9. RESULT USE [X Not applicable]

9.1 Result use is included in evaluation with:

__real users, real use __ real users, controlled use __simulated __no result use_____ [__N __Y, to: _____]

9.2 Is the result use in the experiment representative of the real result use patterns?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The result use is representative of the real result use in terms of, e.g.:

__type of use/search goals __effect on success criteria __effect on information needs

__other:_____ [__N __Y, to: _____]

__the result use of end users is not known/well understood.

THE WHITEBOARD

10. METRICS [__Not applicable]

10.1 Do the metrics measure what matters most to the end users?

__yes __reasonably __no, not very X not known X other: ~~metrics undisclosed as of yet?~~ [__N __Y, to: _____]

↳ the metric is based on binary relations between documents. The weighing is done with a harmonic mean. There is no method to influence that weighing other than through the harmonic mean. (B) is that enough!?

Metrics measure what matters most to the end users in terms of, e.g.:

__cost of errors __efficiency __time spent __effort __domain restrictions

__other:_____ [__N __Y, to: _____]

10.2 Are the metrics used predictive of real world performance?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The metrics are predictive of real world performance, in terms of, e.g.:

__relative performance (between systems) __absolute performance __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

Appendix J. Enterprise search checklists

Background Feature Checklist – Form 1

1. USER ROLE

1.1 Role Name Customer

1.2 Related Roles n/a

	Not known/ not applicable	Related to evaluation
2. INCUMBENTS		
2.1 Domain knowledge: __ none __limited __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: E 1.1]
2.2 General search or system proficiency: __ novice __moderate __expert __varies	[__N/K __N/A]	[__N __Y, to: E 3.1]
2.3 System knowledge: __ none __limited __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: E 1.1]
2.4 Language skills: __ none __limited __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: E 1.1]
2.5 Other relevant user features (e.g., age, training, education, disabilities, etc.): _____	[__N/K __N/A]	[__N __Y, to:_____]
2.6 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]
3. REPOSITORY		
3.1 Media: __text __image __video __ audio __ graphs __3D objects __ varies __other: _____	[__N/K __N/A]	[__N __Y, to: E 4.1, E 6.1]
3.2 Granularity (of what is an information item): __low __ medium __high __varies __specific: _____	[__N/K __N/A]	[__N __Y, to: E 4.1, E 6.1]
3.3 Genre: __news __factual __entertainment __scientific __commercial __personal commentary __technical text __varies __other: _____	[__N/K __N/A]	[__N __Y, to: E 4.1]
3.4 Language: __monolingual __bilingual __multilingual __ other: depends on application and target audience_____	[__N/K __N/A]	[__N __Y, to: E 4.1]
3.5 Technical Quality: __low __moderate __high __varies	[__N/K __N/A]	[__N __Y, to: E 4.1]
3.6 Source Dynamics: __static collection __dynamic collection __stream __other: _____	[__N/K __N/A]	[__N __Y, to: E 4.1]
3.7 Indexing Timeliness: __immediate __every hour __daily __weekly __monthly __varies __other: _____	[__N/K __N/A]	[__N __Y, to: E 4.1]
3.8 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]
4. INFORMATION		
4.1 Origin of user input: __aural __visual __mental __touch __varies __specific:_____	[__N/K __N/A]	[__N __Y, to: E 7.*]
4.2 Clarity of information need: __clear __medium __muddled __varies	[__N/K __N/A]	[__N __Y, to: E 7.*]
4.3 Flow direction: __system to user __user to system __balanced __varies	[__N/K __N/A]	[__N __Y, to: E 7.*]
4.4 Information volume: __low __medium __high __specific: _____	[__N/K __N/A]	[__N __Y, to:_____]
4.5 Complexity of information: __low __medium __high __varies	[__N/K __N/A]	[__N __Y, to:_____]
4.6 Additional features/notes: _____	[__N/K __N/A]	[__N __Y, to:_____]

5. INTERACTION

5.1 Locus of control: __push (system) __pull (user) __varies [__N/K __N/A] [__N __Y, to: E 7.*]

5.2 Complexity of interaction: __low __medium __high __varies __specific: _____ [__N/K __N/A] [__N __Y, to: E 7.*]

5.3 Predictability of interaction: __low __medium __high __specific: _____ [__N/K __N/A] [__N __Y, to: E 7.*]

5.4 Frequency: __rare __recurrent __frequent __varied __specific: _____ [__N/K __N/A] [__N __Y, to: E 7.*]

5.5 Regularity: __irregular__ regular period __varied __specific:_____ [__N/K __N/A] [__N __Y, to: E 7.*]

5.6 Goal-orientation: __random __vague __average __goal oriented __other:_____ [__N/K __N/A] [__N __Y, to: E 7.*]

5.7 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

6. ORIENTATION

6.1 Motivation: __high __middle __low __varies __specific: _____ [__N/K __N/A] [__N __Y, to: E 7.2]

6.2 Likelihood of changing role: __low __medium __high __specific: _____

To what roles and when? _____ [__N/K __N/A] [__N __Y, to: E 7.2]

6.3 Likelihood of abandoning system: __low __medium __high __specific: _____

On what conditions or why? Bad usability, bad responsiveness, lack of content, etc._____ [__N/K __N/A] [__N __Y, to: E 7.2]

6.4 Purpose of use: __Professional __leisure/utility __leisure/entertainment __other: _____ [__N/K __N/A] [__N __Y, to: E 7.2]

6.5 Optionality of use: __ Required use __optional use (conditions): Information is usually available elsewhere, too____ [__N/K __N/A] [__N __Y, to: E 7.2]

6.6 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

7. RESTRICTIONS

7.1 Cost of Errors: __low __medium __high __specific: _____ [__N/K __N/A] [__N __Y, to: E 10.1]

7.2 Time restrictions: __none __low __medium __high __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

7.3 Access restrictions: __none __confidentiality/access rights __pay-per-view __pay-per-search __pay-per-time __other:_____ [__N/K __N/A] [__N __Y, to:E 3.1]

7.4 Device restrictions: __size __processing speed __available other tools or programs __input means _output means __other _____ [__N/K __N/A] [__N __Y, to:_____]

7.5 Network restrictions: __low __medium __high __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

7.6 Additional restrictions and requirements related to organizational context (e.g., coverage requirements, etc.):

_____ [__N/K __N/A] [__N __Y, to:_____]

7.7 Additional features/notes: _____ [__N/K __N/A] [__N __Y, to:_____]

8. PHYSICAL ENVIRONMENT

8.1 Mobility: __mobile __stationary __varies __specific: _____ [__N/K __N/A] [__N __Y, to:_____]

8.2 Geo-position: __one __many __specific:_____ [__N/K __N/A] [__N __Y, to:_____]
8.3 Distractions: __noise __interruptions __parallel tasks __other:_____ [__N/K __N/A] [__N __Y, to:_____]
8.4 Climate and lighting conditions: __lighting __temperature __humidity __other:_____ [__N/K __N/A] [__N __Y, to:_____]
8.5 Additional features /notes:_____ [__N/K __N/A] [__N __Y, to:_____]

9. SUCCESS CRITERIA

9.1 Efficiency: Application must be efficient in terms of presentation and user guidance, as well as performance [__N/K __N/A] [__N __Y, to: E 8.2, E 9.*]
9.2 Effectiveness: Users are expected to be primarily precision-oriented [__N/K __N/A] [__N __Y, to: E 8.2, E 9.*]
9.3 Satisfaction: Based on the ability of a user to find information about the application's company [__N/K __N/A] [__N __Y, to: E 8.2, E 9.*]
9.4 System reliability: Errors lead to immediate abandonment of the system [__N/K __N/A] [__N __Y, to: E 8.2, E 9.*]
9.5 System intuitiveness: Basic web browsing interaction patterns should be supported [__N/K __N/A] [__N __Y, to: E 9.*]
9.6 System comprehensibility: Can help, but generally low priority [__N/K __N/A] [__N __Y, to: E 9.*]
9.7 Actionability: Provided information should support customers in buying decisions [__N/K __N/A] [__N __Y, to: E 8.2, E 9.*]
9.8 Additional criteria: ___ [__N/K __N/A] [__N __Y, to:_____]
9.9 Notes on success criteria:

_____
_____
_____

Form 2 – Interaction and goals

1. USE CASE NAME AND SUPPORTED USER ROLES

1.1 Name: Enterprise Products and Services Search

1.2 Supports (user roles): Customer

2. USER GOALS	Not Know/ Applicable	Related to Evaluation
2.1 Type of information: __single fact/answer/notification __collection of facts/answers/notifications __single item (e.g., document) __collection of items __other:_____	[__N/K __N/A]	[__N __Y, to: E 2.1]
2.2 Type of goal: __viewing __exporting __navigating __ordering/buying (transactional) __manipulating __surfing __other:_____	[__N/K __N/A]	[__N __Y, to: E 2.1]

3. USE CASE RELATIONSHIPS

3.1 Specializes:_____	[__N/K __N/A]	[__N __Y, to:_____]
3.2 Extends:_____	[__N/K __N/A]	[__N __Y, to:_____]
3.3 Uses:_____	[__N/K __N/A]	[__N __Y, to:_____]
3.4 Resembles:_____	[__N/K __N/A]	[__N __Y, to:_____]

4. PATTERN OF INTERACTION – THE USE CASE NARRATIVE

Enterprise Products and Services Search	EXTENDS: **n/a**
USER INTENTION	SYSTEM RESPONSIBILITY
Start interaction Search for a product or service: formulate and enter query optional: Purchase decision, yes/no	 Retrieve and present documents Handle purchase Close (Use Case Ends)
EXTENSIONS	

Form 3 - System and interface feature checklist

	Not Known/ Applicable	Related to Evaluation
1. REQUEST FORMULATION		
1.1 Supported search strategies:		
__querying __browsing __monitoring __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.2 Query persistence: __one shot __permanent __evolving __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.3 Query modality: __text __image __video __audio __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.4 Query formulation: __specification __example __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.5 Query language:		
__simple keyword __basic operators __advanced __specific:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.6 Query target: __content __metadata/description __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.7 Query support:		
__spelling correction __query suggestion __translation __advanced query fields (support for advanced query language __QE __other:_____	[__N/K __N/A]	[__N __Y, to: E 9.1]
1.8 Browsing (content) categories:		
__people __country __subject __media __date __period __language __collection __other:_____	[__N/K __N/A]	[__N __Y, to: E 3.1]
1.9 Navigation support:		
__sitemap __FAQ __classifications __thesauri __other:_____	[__N/K __N/A]	[__N __Y, to: E 7.2]
1.10 Changing between querying and browsing:		
__supported __not supported __specific:_____	[__N/K __N/A]	[__N __Y, to: E 7.2]
1.11 Additional query formulation features/notes:_____	[__N/K __N/A]	[__N __Y, to:_____]
2. RESULT PRESENTATION		
2.1 Presentation hierarchy: __one level __two level __other: varies_____	[__N/K __N/A]	[__N __Y, to: E 6.1]
2.2 Presentation granularity:		
__title __summary __metadata __full information item __set of items __other:_____	[__N/K __N/A]	[__N __Y, to: E 6.1]
2.3 Presentation organization:		
__single item __multiple items __list __ranked list __browsing interface __other:_____	[__N/K __N/A]	[__N __Y, to: E 6.1]
2.4 Result ordering (by):		

__score __date __diversity __author __random __other:_____ [__N/K __N/A] [__N __Y, to: E 6.1]

2.5 Assessment support:

__highlighting __scores __popularity __number of results __relations within a document

__relations between documents __other: _____ [__N/K __N/A] [__N __Y, to: E 6.1]

2.6 Additional result presentation features/notes:_____ [__N/K __N/A] [__N __Y, to:_____]

3. RESULT USE

3.1 Manipulation:

__tagging __annotation __commenting __discussing __creating lists of documents __other:_____ [__N/K __N/A] [__N __Y, to:_____]

3.2 On site consumption/use: __viewing (on screen)__listening (within the system) __analysis and interpretation

__Other:_____ [__N/K __N/A] [__N __Y, to: E 8.*]

3.3 Exporting search context [queries/number of results]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K __N/A] [__N __Y, to: E 8.*]

3.4 Exporting results [single items/sets of documents]:

__saving __printing __publishing (social media, etc.) __other:_____ [__N/K __N/A] [__N __Y, to: E 8.*]

3.5 Sharing: __within the system __exporting __other:_____ [__N/K __N/A] [__N __Y, to: E 8.*]

3.6 Ordering/paying: __internal __external __other: _____ [__N/K __N/A] [__N __Y, to: E 8.*]

Evaluation – Form 4

The feature lists are not meant to be exhaustive, they are just examples meant to help you get started with thinking about how the evaluation task is connected to different use case feature. Please do not let them limit your thinking in any way. Features that do not fit your use case can be skipped. If you think of other features or ideas, write them down on the "whiteboard" under each section.

1. TEST PERSON CHARACTERISTICS [__ Not applicable] **RELEVANT U.C FEATURES**

1.1 Are test persons representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.2.*]

Test persons are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task (e.g. motivation)

__other: End user population is modelled in test scripts, therefore almost completely represented [__N __Y, to: UC 1.2.*]

__ End user population not known/well understood

THE WHITEBOARD

2. TOPICS [__Not applicable]

2.1 Are the topics representative of the real search topics/information needs?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 2.2.*]

Topics representative of the search topics/real information needs in terms of, e.g.:

__domain of topics __clarity of information need (clear/muddled): __type of search goal

__information need durability.

__other:_____ [__N __Y, to: UC 2.2.*]

__Real information needs/search topics not known/well understood

2.2 The topics are used for

__relevance assessments __automatic runs __tasks given to test persons

__other:_____ [__N __Y, to:_____]

THE WHITEBOARD

3. REQUESTS [__Not applicable]

3.1 Are the requests representative of the real requests?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.2.2, UC 3.1.1-6, UC 3.1.8, UC 1.7.3]

The requests are representative of the real requests in terms of, e.g.:

__type of request __request modality __query length __query quality __query structure

__query formulation (specification/example) __query durability

__other:_____ [__N __Y, to: UC 1.2.2, UC 1.2.3, UC 3.1.1-6, UC 3.1.8]

__Real requests not known

THE WHITEBOARD

4. DATA [__Not applicable]

4.1 Is the data used in the evaluation activity representative of the real data?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.3.*]

The test data used is representative of the real data in terms of e.g.:

__intellectual content __modality __size __dynamics __curation __provenance (created by one/created by many)

__structure __granularity

__other:_____ [__N __Y, to: UC 1.3.*]

__Real data not known [__N __Y, to:_____]

THE WHITEBOARD

5. GROUND TRUTH CREATION [__Not applicable]

5.1 Ground truth captures:

__relevance of documents to topics __which of two documents is more relevant to a topic
__which of two ranked lists is better __other:_____ [__N __Y, to:_____]

5.2 Ground truth is obtained:

__ manually __semi-automatically (e.g. pooling) __fully automatically __other:_____ [__N __Y, to: _____]

5.3 Are the relevance criteria representative of the real users' relevance criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Relevance criteria are representative of the real users' relevance criteria in terms of e.g.:

__strictness of criteria __type of criteria (e.g. topicality or novelty) __grades of relevance
__other:_____ [__N __Y, to:_____]
__Real users' relevance criteria not known/well understood

5.4 Are assessors representative of the end user population?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

Assessors are representative of the end user population in terms of, e.g.:

__demographics __search skills __domain knowledge __language skills __relation to search task
__other:_____ [__N __Y, to:_____]
__ End user population not known/well understood

5.5 Are the results shown to assessors representative of the real results shown to end users?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

THE WHITEBOARD

6. RESULT PRESENTATION [__Not applicable]

6.1 Is the result presentation in experiment representative of target system(s)?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.3.1, UC 3.2.*]

The result presentation is representative of target system(s) in terms of e.g.:

__presentation hierarchy __granularity __other:_____ [__N __Y, to: UC 1.3.2, UC 3.2.*]

__Target system result presentation not known

THE WHITEBOARD

7. INTERACTION [__Not applicable]

7.1 Interaction in the experiment is:

__real user interaction __interaction model __other:_____ [__N __Y, to: UC 1.4.1-3, UC 1.5.*]

7.2 Is the interaction in the experiment representative of real end user-system interaction?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.4.1-3, UC 1.5.*, UC 3.1.9-10]

The interaction is representative of real end user-system interaction in terms of e.g.:

__search strategies __result assessment __goal orientation __learning

__session length/complexity __query reformulation (strategies, cost, …)

__other:_____ [__N __Y, to: UC 1.5.6, UC 1.6.*]

__Real interaction patterns not known/well understood

THE WHITEBOARD

8. RESULT USE [__Not applicable]

8.1 Result use is included in evaluation with:

__real users, real use __ real users, controlled use __simulated __no result use_____ [__N __Y, to: UC 3.3.*]

8.2 Is the result use in the experiment representative of the real result use patterns?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 3.3.*]

The result use is representative of the real result use in terms of, e.g.:

__type of use/search goals __effect on success criteria __effect on information needs

__other:_____ [__N __Y, to: UC 1.9.*, UC 3.3.*]

__the result use of end users is not known/well understood.

THE WHITEBOARD

9. EVALUATION CRITERIA [__Not applicable]

9.1 Are the success criteria in the experiment representative of end users' success criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.9.*]

The success criteria are representative of end users success criteria in terms of, e.g.:

__volume of relevant results __time spent (if the goal is to spend time) __user satisfaction

__meeting user expectations __task completion __objectivity/subjectivity of criteria

__other:_____ [__N __Y, to: UC 1.9.*, UC 3.1.7]

__End users' success criteria not known/well-understood

__Evaluation is not based on user criteria, but:_____

9.2 Are the failure criteria in the experiment representative of end users' failure criteria?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: UC 1.9.*]

The failure criteria are representative of end users' failure criteria in terms of, e.g.:

__time __effort __poor result quality __frustration __"out of queries" __other: _____ [__N __Y, to: UC 1.9.*]

__End users' failure criteria not known/well understood

10. METRICS [__Not applicable]

10.1 Do the metrics measure what matters most to the end users?

__yes __reasonably __no, not very __ that's not the goal __not known

__other:_____ [__N __Y, to: _____]

Metrics measure what matters most to the end users in terms of, e.g.:

__task completion __cost of errors __efficiency __time spent __effort __domain restrictions

__other:_____ [__N __Y, to: UC 1.7.1, UC 1.9.*]

10.2 Are the metrics used predictive of real world performance?

__yes __reasonably __no, not very __not known __other:_____ [__N __Y, to: _____]

The metrics are predictive of real world performance, in terms of, e.g.:

__relative performance (between systems) __absolute performance __other:_____ [__N __Y, to: _____]

THE WHITEBOARD