Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on results of the WP3 first evaluation phase

Deliverable number	*D3.2*
Dissemination level	*Public*
Delivery date	*31 August 2012*
Status	*Final version (Aug 30th, 2012)*
Author(s)	*Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Lorraine Goeuriot, Jessica Ignalski, Matthias Jordan, Liadh Kelly, Sascha Kriewel*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Table of Contents

Executive Summary

This deliverable presents the report on the first interface component evaluation phase for the Khresmoi search system. It describes a number of formative usability and learn-ability evaluations that were conducted, some using an eye-tracking system, as well as experiments comparing different variants of interface components for effectiveness and efficiency. Three usability evaluations revealed a number of possible areas for improvement, but also confirmed the general suitability of the user interface framework and the search tool. The collaborative features underwent a first user test and while the implementation clearly showed its early stage of development, the concept including a personal library, a user profile and the possibility to share results with other users is sound.

Component evaluations that compared different ways of presenting suggestions or faceted result lists were inconclusive and they seem mostly a matter of preference. They will receive further refinement and additional evaluations in the future. An evaluation of different result item presentations clearly showed significant differences in their suitability for judging the relevance of search results. Two new result presentations performed significantly better than the ezDL baseline when used by searchers. Also described are the logging framework and the evaluation setup in the form of user evaluation scripts and questionnaires that will be used to conduct future evaluations.

List of abbreviations

AOI	*Area of Interest*
ANOVA	*Analysis Of Variance (a statistical test)*
CLEF	*Conference and Labs of the Evaluation Forum*
ezDL	*easy Access to Digital Libraries (an open source user interface framework)*
RDF	*Resource Description Framework*
RIA	*Rich Internet Application*
ROI	*Region of Interest (see Area of Interest)*
RSV	*Retrieval Status Value*
QUIS	*Questionnaire for User Interface Satisfaction*
SMI	*SensoMotoric Instruments GmbH (producer of eye tracking software and equipment)*
SUS	*System Usability Scale*
UI	*User interface*
UX	*User experience*

1 Introduction

The evaluation of user interface components generally falls into two categories: evaluations of the usability of the user interface and (comparative) evaluations of the efficiency of specific components for performing specific, well defined tasks. Usability is an umbrella term that comprises a number of features which make user interfaces user friendly. First off, it should be easy to learn and use without prior knowledge – unless it is a specialized interface for experts who are expected to spend a lot of time working with the interface. In that case the second aspect of usability comes into the forefront: users should be able to quickly and efficiently perform tasks. Usability also means that it should be easy for users to recover from errors (or even avoid errors in the first case), and finally it should be pleasant to use. The last part of usability is often summarized under the term "user experience" (UX), the highly subjective, affective and emotional aspects of working with a user interface.

At this point during the project, many components are still in prototypical form. A comparative evaluation with existing user interfaces is therefore not likely to provide any insights. Instead, a formative evaluation is called for, in which possible variants of interface components are compared and tested for usability and learn-ability using established measurement tools such as usability questionnaires.

This deliverable describes a number of experiments conducted to test the usability of specific features of the Khresmoi search interface (based on the ezDL framework), as well as three comparative evaluations designed to find the most appropriate variant for particular tasks. During the usability experiments which are described in Sections 3.2.1, 3.3.1 and 3.4, participants were asked to perform a number of common tasks without any introduction or explanation of the search system. Through observation, logging, recording of user interaction, usability questionnaires and interviews problems with the usability of the features can be discovered.

Three other experiments (detailed in Sections 3.1, 3.2.3 and 3.2.4) compared different variations of UI features for their suitability in effectively performing specific tasks: using query term suggestions and making quick relevance judgements for lists of search results.

Section 2 starts with describing the evaluation infrastructure implemented for conducting these experiments. It can and will be used for further experiments in the upcoming months. Section 3 then presents the evaluations conducted and possible further experiments based on the results described in this deliverable are detailed in Section 4.

2 Evaluation and Logging Infrastructure

The evaluation system used during the experiments is based on the ezDL system described in [1]. A detailed description of the system and the extensions developed during the Khresmoi project will be given in deliverable D3.3. To support user-centred evaluations, the ezDL framework has been extended with an evaluation mode that addresses many of the major challenges inherent in setting up evaluation tasks and tracking user activity during the experiments:

- Controlling the experimental setup (e.g. presenting search tasks to the user, rotating search tasks)
- Logging the visibility data of the components on the screen to create *area of interest* (AOI) data [19]

- Logging the actions of the user

Tasks can be defined and structured using a task tool (pictured in Figure 1) and a special starter (Figure 2). The starter allows for selecting a specific experiment setup, the task tool controls the experiment flow and makes it easy to use common rotation patterns, such as Latin square or Graeco-Latin square designs.

For recording AOI data the AOILog framework [2] was integrated into ezDL.

The following gives a brief overview of the logging mechanism developed in WP3 to assist in user evaluations. It was used for this first round of evaluations and will also find use in the evaluations planned for WP10. During the evaluations described in this deliverable a simple storage mechanism using a local relational database was used to store the logging data. In the future, the integrated Khresmoi prototype will be able to store the logging data in an RDF store.

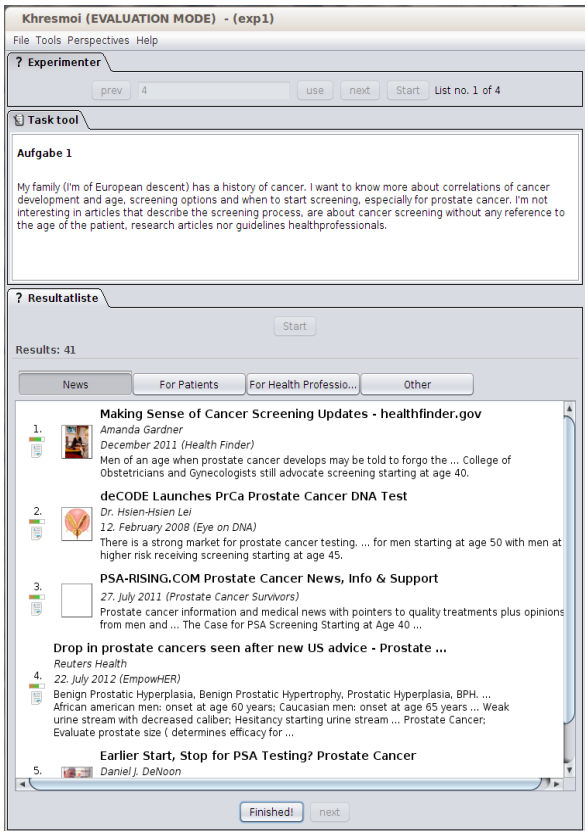


Figure 1: The task tool for user experiments.

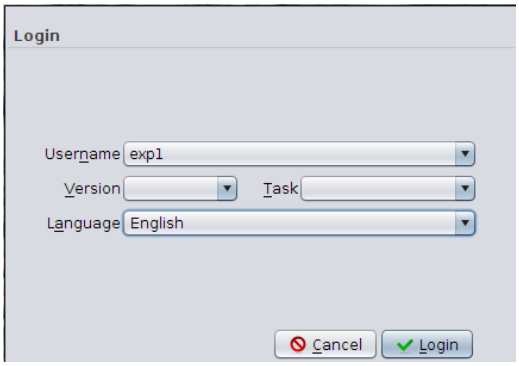


Figure 2: The login screen for user experiments with task chooser.

2.1 Logging user actions

For user evaluations all actions performed with the system should be logged for later inspection and analysis. The ezDL framework was extended by a **`UserLogManager`** that can easily be switched off for production use and in switched-of mode has little to no effect on system performance. Log events are anonymous classes that implement a comment interface **`UserLogEventProvider`** and provides methods for retrieving a log event name and log event parameters. All log events generated in the front-end are sent by the manager to a **`UserLogAgent`** in the ezDL back-end, which currently stores all the interaction data of the user in a relational database (currently MySQL is used). A log session is comprised of all log events that a user or the system has triggered. A log event has

- a unique name identifying this type of event (defined in `UserLogConstants`),
- timestamps from the front-end and the back-end,
- a sequence number to ensure the correct order, and
- parameters as multiple key/value pairs.

A complete list of current log events is given in Appendix 7.6. For example, when a user performs a search for "diabetes mellitus" within the Textual- and the ImageSearch services of Khresmoi the corresponding log event might look like this (the sequence number indicates that the event was the 10th event during this user session):

name: "searchstart"
clientTimestamp: 1/4/2012 15:26:32, 1234
timestamp: 1/4/2012 15:26:32, 3456
sequenceNumber: 10
parameters:
query: "diabetes mellitus"
sources: "textsearch, imagesearch"

The logging service of ezDL takes care of allocating activities to sessions and users. During the following experiments each participant was assigned a unique user account, making it easy to identify the events of a particular evaluation session. Should later evaluations require logging of new actions these can be simply integrated by sending a corresponding logging message to the back-end.

2.2 Future RDF logging

In the future, the logging data will be stored in an RDF-based ontology of the user model. The user model in Khresmoi consist of three parts:

- **User profile:** Defines the unique characteristics that distinctively identify and describe a given user, such as their name and email address, their preferred language or search topics, or their level of medical knowledge. It could possibly also include a health profile if the user is willing to provide such to the system.
- **User session:** Defines attributes of the user's current search session, such as the type of connection, screen resolution, chosen perspective, device type and operating system, system language and current location.
- **User activities:** Describe the different types of actions, queries and search activities supported by the system. A user session is composed of a series of user activities.

The log events in ezDL will be mapped to RDF triples that will then be stored in the corresponding user model using an RDF ontology store. User activities are based on previous work on a log scheme for digital library evaluation [17, 18] and represent single actions of users within a longer user session. Within this logging scheme, specific user activities and tool uses are mapped onto more general conceptual activity types which can be used for high level analysis. For example, all user actions that involve storing objects for later use (by printing, exporting to a local folder, mailing to an external address, copying to a clipboard or saving to the personal library), are considered instances of the Store activity and it is thus possible to abstract from the concrete location and method of storing the object. An overview of the logging scheme can be seen in Figure 3.

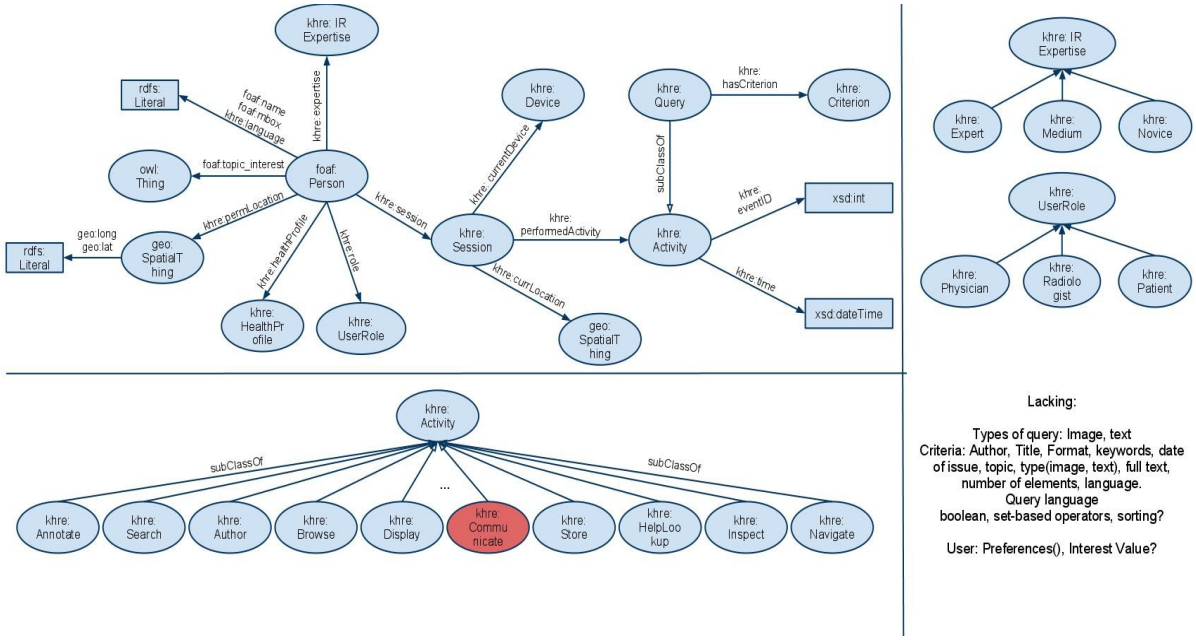


Figure 3: Ontology for logging user activities.

3 Experiments Conducted and Results

The first component evaluation phase concentrated on a formative evaluation of the user interface features. For most of the interface components only a single version was developed. It is therefore "not possible to demonstrate how much better users perform" compared to some other component [1]. Usability tests are the most appropriate type of evaluation at this point during the development. A summative evaluation as well as more comparative evaluations will be conducted during the second component evaluation phase, to measure the performance of the finished features and test the final system.

The user interface components that were developed during the first development phase of the Khresmoi project fall mainly into the areas of the flexible interface tool-kit (task T3.1) and the query specification and result presentation (task T3.2). Following the results of the use case requirements analysis, the development of collaborative components was added to the work plan, and a first evaluation was included in this evaluation phase.

The component evaluations done in WP3 were conducted independently of the integrated application. While progress for some of the objectives of WP3 can (only) be measured in an integrated system as part of the evaluation strategy for WP10, a number of objectives can be evaluated isolated from the larger system. In particular, the user-centred evaluations for WP3 try to determine if specific interface features are task appropriate and easy to use or learn. Therefore, the various components (presentation of suggestions, support of search and result manipulation, result presentation, and support collaboration) were treated as isolated features, which could then be tested in a lab setting for functionality, usability, learn-ability, subjective usefulness and user acceptance. The results will be used to refine the existing components or choose between different interface concepts (e.g., for presenting search suggestions, single results, or a categorized result list). This is expected to be an iterative process, which will finally lead to mature and usable interface components that are suited for the specific tasks required by Khresmoi users.

The following evaluations were conducted:

- **Evaluation of query suggestion components:** Within task T3.2 query refinement and suggestion components are being developed. These present query formulation support using sources from work packages four and six (e.g., spelling correction, disambiguation of terms, suggestions of related terms or related queries). The user-interface part of the search refinement or query reformulation and expansion options consists in presenting these options to the user. The evaluation described in Section 3.1 tried to assess how different presentations allow searchers to make use of the suggestions presented.
- **Evaluation of result presentation components:** Within task T3.2 several result presentation components were developed. These present a result list or parts of a result list interactively and allow for filtering, sorting, grouping, extracting of the results and faceted navigation. The result surrogates used should be able to display all information necessary to support the use case requirement collected in deliverable D8.2. The presentation should allow users to efficiently and effectively assess the relevance and suitability of results, thus preventing them from spending unnecessary effort on examining the details of non-relevant results. Three separate experiments were conducted during the first evaluation phase: a usability test of the search tool including the result interaction components (detailed in Section 3.2.1), a comparative evaluation of different presentation variants for result items (in Section 3.2.3), and a comparative evaluation of two variant for presentation and interacting with result categories (in Section 3.2.4).
- **Evaluation of collaborative components:** As part of the interface framework developed in task T3.1, a personal folder for storing favourite search results was provided. This folder allows users to collect documents (including documents uploaded by the users themselves) and share them with other users. Within these folders, users can add tags or labels to documents. They can also manage a personal profile and search for other users' profiles. The evaluation described in Section 3.3 is a first test of the usability and learn-ability of these functions.
- **Evaluation of flexible framework for web search system:** Within task T3.1, an alternative interface framework to the comprehensive, stand-alone Java client was developed. This interface is a Rich Internet Application (RIA) that can run in an ordinary web browser, yet it provides many of the interaction and customization options of the stand-alone interface. The experiment described in Section 3.4 was an evaluation of the usability and learn-ability of those interface features.

If not otherwise mentioned, the experiments were conducted at the usability lab of the Information Engineering group of the University Duisburg-Essen. Participants were compensated with certificates (for students who could use them in exchange for course credit) or with €5 per half hour of participation. A common evaluation script was designed based on recommendations and examples given by Snyder [3]. This script was followed by all experiment facilitators (the script is reproduced in Appendix 7.1). All participants received an explanation of the experiment and signed a consent form releasing the data collected for use and analysis within the project. For some of the experiments conducted at the University of Duisburg-Essen, an eye-tracking system (the RED eye-tracker by SMI) and a web cam was used. The screen-recordings, gaze recording, and videos collected were analysed and anonymised results from these analyses have been included in this report.

3.1 Evaluation of Query Specification Components

The goal of this first evaluation was to determine which of two variants of the proactive query specification support component best supports the user. Design decisions to be evaluated were how the query suggestion pop-up should be presented (as a list with comments or organised in tabs) and how to notify the user that suggestions are available for a certain part of the query. A more detailed description

of the evaluation was published in [14].

3.1.1 Pre-tests of Mock-ups

A pre-test was conducted to see which design mock-up users would like best with respect to triggering the suggestions and presenting them. Three design variants were examined: a simple list, as known from many web search engines (Figure 4), a coloured list (Figure 5), and a variant using multiple windows (Figure 6).

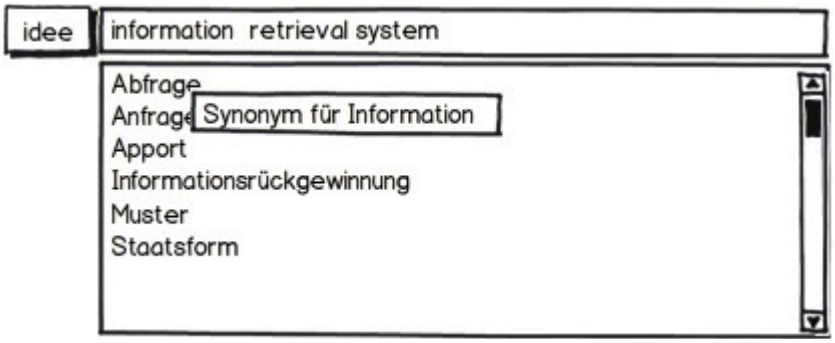


Figure 4: Mock-up for simple suggestion list.

The simple list variant was designed to open a drop-down list with alphabetically sorted suggestions each of which had a description of the kind of suggestion (e.g. "synonym") to be opened by a mouse-over.

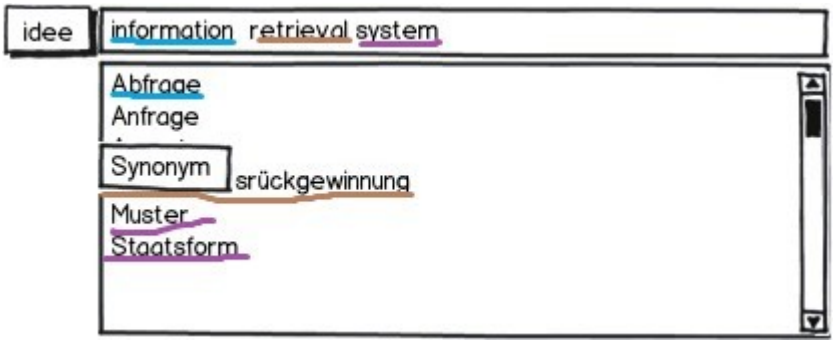


Figure 5: Mock-up for colour-coded suggestion list.

The idea behind the coloured list was to underline each term in the query for which suggestions can be found in the list. The connection between the suggestion and the term for which the suggestion is intended is visualized by underlining the suggestions with the same colour that is used to underline the term. This variant also shows a description for the suggestion on mouse-over.

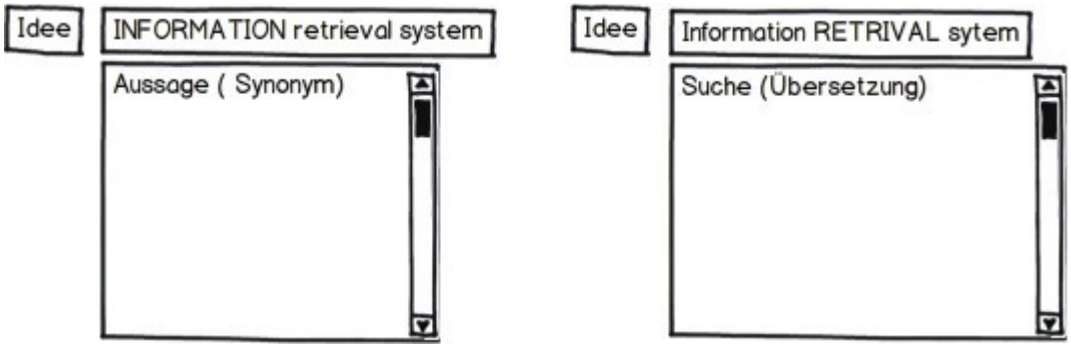


Figure 6: Mock-up for suggestions in multiple windows.

The third variant opened a separate list for the suggestions for each term. All variants work on term basis. Using a single concept – opposed to also offering suggestions on query basis – allows comparison of the visualizations.

The mock-ups were evaluated by presenting the designs as a paper print-out to each participant along with a description of each design. The experimenter answered questions and conducted an interview about the participant's thoughts about the designs. The benefits of paper mock-ups for early UI testing are detailed in [3].

Participants were recruited using a standardized text in web forums of two student bodies at the University of Duisburg-Essen. The participants were between 21 and 32 years old (mean = 24 years) and selected to be experienced and practised in using search engines. The analysis of the data was done according to Mayring [6].

The results showed that the simple list and the coloured list were not favoured by the participants. A common criticism was that the alphabetical sorting required too much cognitive effort for evaluating the list. The mouse-overs with descriptions were not well-received because mouse-overs took too long time to open and had to be opened for each item. The third design was received better; positive comments included that it was obvious which term the suggestions belonged to.

Based on these results, the third design, which showed suggestions for each term in a separate window to be opened individually, was used for the following evaluations. As an alternative to this design a second design was evaluated that showed suggestions in separate tabs – one for each category the suggestions belonged to (e.g. synonyms, spelling corrections, translations). This second design was introduced because of the participants strong rejection of two mock ups to allow a comparative evaluation of the following experiments.

3.1.2 Experiment A1: Position of Hint Icons

As a way of notifying the users of available suggestions for some terms in a text field, a light bulb icon was introduced. This idea is based on an observation by Schaefer et al. [5] who found that users tend to miss the suggestion pop-ups if they are only shown for a short moment. The light bulb icon was supposed to be lit if suggestions can be shown, regardless of whether the list was actually shown or not. Experiment A1 was conducted to find the optimal place for these icons.

Three different positions were tested: to the left of each query text field, to the right of each query text field, and one light bulb icon placed next to the “search” button, indicating available suggestions for any of the query text fields (see Figure 7).

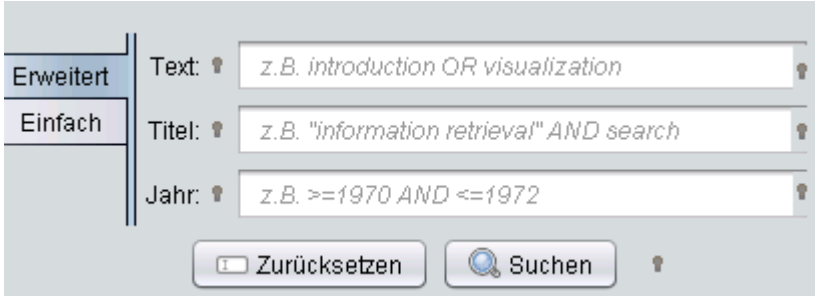


Figure 7: Positions of the "light bulb" hint icon.

Using the CLEF collection, participants performed search tasks using one of the three different designs. The experiment was recorded using the eye-tracker RED by SMI.

If the system could offer suggestions to any query term, the light bulb icon(s) were lit and the terms with suggestions were underlined in blue. Users could then open the suggestions by clicking on the underlined terms.

The participants were recruited again using a standardized text in forums of the University of Duisburg-Essen. The recruited 9 participants were students of the university, high-school students or employed persons in the age between 22 and 24 years of age; four were male, five were female. All stated that they frequently use search systems and favour Google. They were assigned randomly one of the three experimental conditions, so that to each condition three participants were assigned.

Each experiment consisted of three tasks from the CLEF collection the participants were asked to perform. After the experiment an interview was conducted along five main questions: “how did you like the handling of the system?,” “what did you like?,” “why?,” “what did not you like?,” and “why not?.”

The eye-tracker data was evaluated with respect to three so-called areas of interest – rectangular areas of screen estate around each of the light bulb icon groups. The evaluation showed that the participants using the icons at the left side of the query form fields looked eight times at those icons on average. Participants in the group with the icons placed at the right of the form fields noticed the icons once on average. The icon next to the search button was not noticed by any of the participants using this interface. A Kruskal-Wallis test on the data gave a $p = 0.023$, being significant at the 0.05 significance level. A post-hoc test using the Mann-Whitney U resulted in $p = 0.1$ for each pair of groups, making it impossible to determine the root of the difference reported by the Kruskal-Wallis test.

The interviews found that all participants found the ezDL interface not more difficult than the Google interface. Some participants reported they felt they needed some time to get used to the interface. Interestingly, only four participants reported in the interview to have noticed the light bulb icons at all.

Because of these results, duplicate visualisation of suggestion availability and the idea of using the light bulb icons were dropped. Instead, further interfaces only use coloured underlines to notify the user of suggestions and clicking on the marked terms opens a drop down list with suggestions for that term.

3.1.3 Experiment A2: Presentation of Suggestions

The last experiment in the campaign was conducted to test the difference between presenting suggestions in one single list pop-up and presenting them in multiple lists in different tabs of a pop-up (see Figure 8). Variables measured were acceptance rate for the suggestions, learn-ability, user feedback, usability and how often users noticed the suggestions.

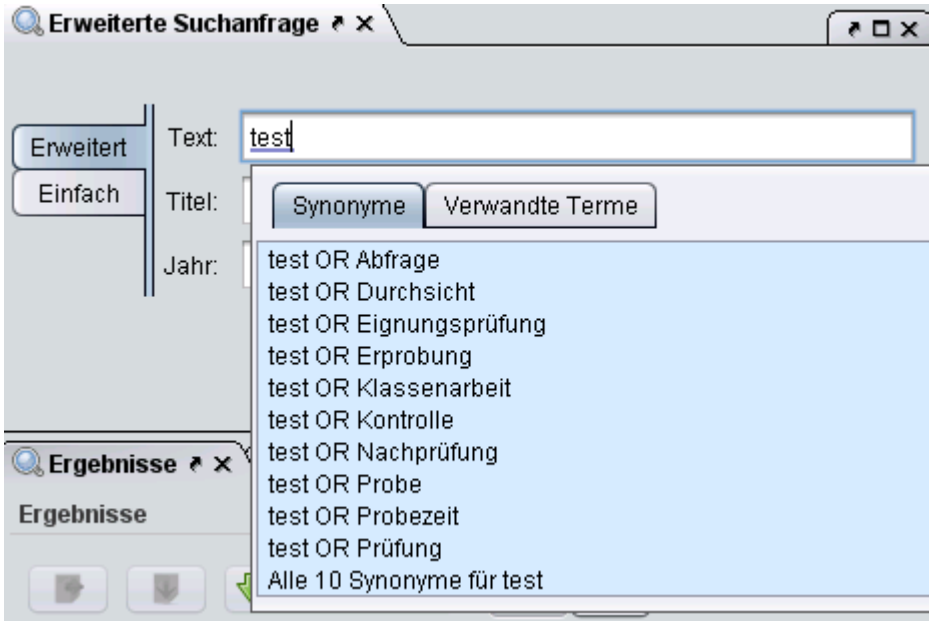


Figure 8: Suggestion dialogue with tabs.

To open the suggestion pop-up (regardless of design), the user could either pause during typing or click on a marked term.

The study involved 18 participants, who were recruited using a standardized text in forums of the University of Duisburg-Essen and on flyers distributed on campus. Participants were between 19 and 29 years old with a mean of 22.2 years of age (SD = 2.73). Among the participants were students of the university as well as non-students. Six participants were male, 12 female. The participants had between 6 and 14 years of experience with search engines (mean = 8.89 years, SD = 2.32). Sixteen participants were very satisfied with their previous experiences with search systems, one participant was only somewhat, and one participant was not satisfied at all.

After instructing participants about the interface and calibrating the eye-tracker, they had to perform three tasks with a time limit of 15 minutes each. Afterwards each participant had to complete a QUIS (Questionnaire of User Interface Satisfaction, see [11]), an additional questionnaire about their assessment of the search process and open questions regarding the suggestion drop-downs.

The analysis of the recorded data showed that participants using the tabbed suggestion drop-down clicked more often on the suggestions than those in the other experimental condition. The average number of fixations on the drop-down lists was 94.56 (SD = 78.4). Since the occasions during which suggestions were available to the participants at all were different between the participants, a variable "click rate" was calculated by normalizing the number of actual clicks to the number of possible clicks. The click rate for the list-based suggestions was 0.08 and that for the tabbed display was 0.06 but a Mann-Whitney U showed no significance at any significance level.

The learn-ability was measured by calculating a score from answers to various questions about the respective designs. Of the 18 participants 7 answered all questions correctly (38.9%). A t-test, performed after confirming normal distribution of the data, using the Kolmogorov-Smirnov test, and equal variances, using the Levene test, gave $p = 0.496$, indicating that it is unlikely that there is an actual difference in learn-ability between the designs. The difference between the average scores of each design was 0.41, which seems large given a 1.7 average score for the list version. Considering the standard deviation of 0.36 for that variant puts the score difference into perspective.

The subjective assessment of the search process was measured using five items in the questionnaire. A Mann-Whitney U between the scores for each design gave a $p = 0.796$, which is far off the range

where it would seem plausible to assume a difference between the designs concerning the assessment of the search process.

The fourth question was related to the difference regarding usability between the designs. Usability was operationalised as a mix of user satisfaction, efficiency and effectiveness. Our effectiveness measure was the F-measure, the harmonic mean between recall and precision. The efficiency measure was effectiveness per time. The satisfaction of the user was estimated using the QUIS. The measures were subjected to Mann-Whitney U significance tests, none of which gave values for p anywhere near the 0.05 significance level.

The question how often users noticed the suggestions, was answered by using the eye-tracker data. The number of fixations on the list variant was 122.7 on average compared to 98.6 fixations on the tabbed variant on average, but the standard deviation was 33.9 for the list version, being larger than the difference of fixation averages. As expected, a Mann-Whitney U gave a $p = 0.730$, being clearly in the range where the hypothesis that there is a difference between both designs with respect to the number of fixations has to be rejected.

Unfortunately the results do not show a clearly superior design choice for presenting suggestions with regards to the effectiveness in supporting users or in obvious usability measures. It seems mainly a question of preference. For the Khresmoi search system both variants will be offered as plug-ins making it a choice of system maintainer or even the individual user.

3.2 Evaluation of Result Presentation Components

Within task T3.2 several result presentation components are being developed. These present a result list or parts of a result list interactively and allow for filtering, sorting, grouping, extracting of the results and faceted navigation. The result surrogates used should be able to display all information necessary for supporting the use cases and user stories. The presentation should allow users of the system to more efficiently and effectively assess the relevance, usefulness and suitability of results, thus preventing them from spending unnecessary effort on examining the details of non-relevant results.

For the evaluation of the result presentation components three experiments have been conducted. Experiment B1 measures the usability of the search tool. Experiment B2 compares different result item presentations by using a questionnaire and eye-tracking. Experiment B3 investigates whether the users prefer a grouped or tabbed result item presentation.

3.2.1 Experiment B1: Usability Tests of Search Tool

During the month of July 2012, a usability evaluation of the search and result interaction for the Khresmoi interface was conducted. This evaluation took place at the University of Duisburg-Essen and Dublin City University. Experiments at the University of Duisburg-Essen were recorded with eye-tracking equipment (screen-recordings, mouse-movements, and gaze data). Complete logging data of the user interactions was collected for all experiments. In addition, one or two facilitators kept notes during the experiments and performed a short interview with the participants.

During the experiments, a total of 8 participants were asked to perform a number of typical tasks with the user interface that covered different ways of searching for documents, filtering, grouping and sorting results, extracting keywords, looking at the result preview and opening the original web page, opening full-size images from thumbnails and performing a search for similar images, switching between result list variants, and working with the query history.

The participants were given no introduction to the user interface and did not receive help during their tasks (unless they were stuck and could not continue). All participants were initially unfamiliar with

ezDL or the Khresmoi interface. The goal of the experiments was to discover usability and learnability problems that need to be addressed. A complete description of the tasks for this evaluation is given in Appendix 7.2.

Three of the 8 participants were female and 5 were students at various stages of their studies (graduate and undergraduate students). Four used the German version of the interface, with the remaining participants using the English version. Before the actual tasks, the participants were asked about their previous experience with computers and web search in general, but also about their experience with systems that provide similar functionalities as those under consideration. All participants used computers regularly or often. Seven of the participants were regular users of web search engines, while one only rarely used a computer to search the web. All had previously used the internet to search for health related issues, but only one did so often.

Not all participants were able to successfully complete all the tasks. In particular, many had problems opening a view of their previous queries ("search history"). This was discovered to be due to a problem in the docking framework used to arrange the tool views within the interface. Also related to the docking framework was the problem of two users who accidentally closed important parts of the search interface (the query box in one case, the result list in the other) and were not able to recover from this situation on their own.

On average, users needed 13 minutes 47 seconds to complete the experiment (between 6 minutes and 25 minutes 34 seconds). The longest task was *task 6* ("Search for the term that most frequently occurs within the results") with 3 minutes 26 seconds on average. It was clearly the most unfamiliar task to the searchers and even the two fastest participants needed about a minute to complete it. A heat map of one user who is trying to find the extraction is shown in Figure 20 (Appendix 7.7). While extracting terms from result sets takes some processing time, it would be worth considering to do this extraction automatically and display a term cloud visualisation of the most common term alongside the initial result list (instead of having the user initiate the display). In addition the labels for some of the interactive element seems to have been confusing to the users, who did not know what to expect.

Some of the fastest tasks clearly were not raising problems for any of the users, such as resizing a thumbnail image (by clicking on it), opening the full text in a browser (by following a link labelled "detail link" and marked with a special icon) and searching for similar images (by following a link labelled "similar images"). Setting a filter on the search result was also accomplished by all participants, although some took longer than others to find the filter box (between 17 and 75 seconds). Removing the filter was easily accomplished by all (in 2 to 30 seconds), but few participants used the "clear icon" to do so, instead most participants deleted the text character by character or selected and then deleted the filter text.

The task with the widest variation was *task 12* ("switch from result grid to result list"). While three users were immediately able to find the toggle button (needing between 9 and 18 seconds to locate and click it), users #2 and #3 both needed over two minutes. The icon and interaction used is quite common in other search systems and applications, and the iconography clearly shows the two types of result display. However, as one participant remarked, the toggle was missing a tool-tip which might have been confusing for someone who was not already familiar with this UI element. A heat map and a scan path of one user who is trying to find the toggle button are shown in Figure 21 and Figure 22(Appendix 7.7).

The following is an example for a timestamped task completion protocol of an evaluation session:

User #2: total time 14:30	
0:21	Cursor placed in search box

0:26	Search for “aspirin” finished (task 1 completed)
1:00	Sorted by year descending (task 2 completed)
1:39	Set filter for “asthma” (task 3 completed)
1:54	Removed filter (task 4 completed)
2:04	Cursor placed in search box
2:12	Search for “health diet” finished (task 5 completed)
3:46	Extracted a term cloud of frequent terms
4:10	Used a term to filter (task 6 partially completed)
5:32	Opened first detail
6:01	Seen three details (task 7 completed)
6:11	returned to first document (task 8 completed)
6:30	Opened full text in browser (task 9 completed)
6:55	Returned to program
7:07	Opened full size version of image thumbnail (task 10 completed)
7:39	Search for similar images (task 11 completed)
9:50	Switched back to original result display (task 12 completed)
10:30	Undocked search box
11:09	Opened query history
11:14	Repeated first query (task 13 completed)
11:34	Grouped results by decade
14:30	Found the number of results per decade (task 14 completed)

In addition to the time protocol, a problem protocol was created by the experiment facilitator and observer. After the experiments five developers reviewed the video material and expanded the problem protocol from their observations. A number of usability problems were discovered, which are currently being addressed or have already been addressed within tasks T3.1 and T3.2:

- Participant expects a context menu for item in the detail view.
- New tools opened by users are initially mostly hidden and therefore hard to discover.
- Participants were not able to rearrange the views according to their needs (Some hints were given without result).
- Participants were not able to open the query history view as expected.
- The filtering and grouping functionalities were mostly used without any problems.
- Participants often did not understand that “ghosted” entries in the menu mean that the tool is already open.
- Participants frequently use the tool tips and seem confused by control elements not having one.

- The back/forward mechanism for browsing through open detail views was never discovered.
- The link for opening the full document could be more clearly labelled, possibly as a prominent button (maybe as part of the button bar at the top, as suggested by one participant).
- The "extraction" of terms was not understood by some participants.
- Loading of image documents takes place in a reversed order.
- One participant accidentally undocked a view from the main window and got confused. He was not able to recover from the situation without guidance by the examiner.
- Participants made use of the scroll bar more often than expected. The mouse wheel was hardly ever used. If asked, participants replied that using the scroll bar gave them more control over scroll speed.
- One user accidentally closed the result list and was not able to recover from the situation.
- For one participant the query history obscured the detail view for results.
- The full functionality of the extraction view was not discovered by any of the participants. It was not clear that clicking a term triggers a search for that term. Instead participants typed the terms they found with the tool.
- The wording of the context menu functionalities could be improved as was stated by some participants.
- A few participants made use of the extraction tool in a creative manner by using it for finding the decade with the most publications regarding one topic.
- To some participants it was unclear that extraction always takes place based on the current result set, they assumed it works on the current selection instead.
- One participant suggested that the focus of the mouse wheel should always be on the component currently pointed at.

After the experiments the participants were asked to fill out a usability questionnaire (the SUS questionnaire [7], which was extended by additional questions from the QUIS questionnaire). The complete English questionnaires are listed in Appendix 7.5. The questionnaires were translated into German for the experiments conducted at the University of Duisburg-Essen.

Results from the questionnaires are summarized in Figure 9 and Figure 10 with each question item as one horizontal bar. Figure 6 shows the results for the SUS questionnaire with the negatively formulated items (2, 4, 6, 8, 10) re-coded. In both figures, the light and dark green parts of the bars stand for lightly or strongly positive answers, yellow for a neutral answer, orange and red for lightly or strongly negative answers. The length of the bar segments show the numbers of participants that gave this answer.

For the SUS, questions 1, 3, 7, 9 and 10 received any negative replies, with question 9 (*I felt very confident using the system*) being the most negative. It is perhaps not very surprising that using an unfamiliar system with many new functionalities felt overwhelming to the participants. However, they did not feel that they would need the help of an expert to understand it (question 4) and most believed that users would be able to quickly learn to use the system on their own (question 7).

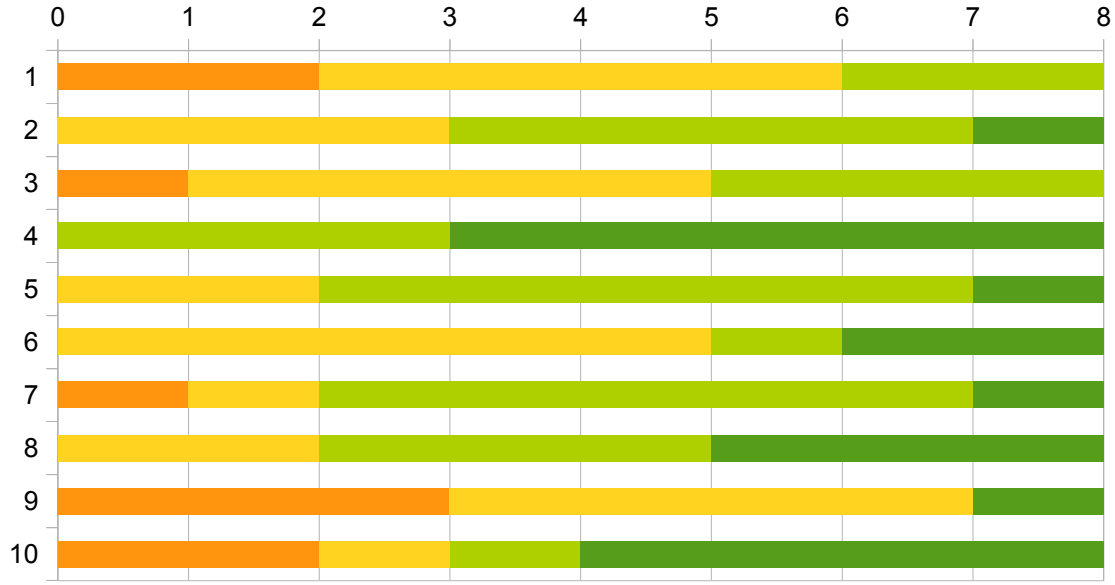


Figure 9: Results of the SUS questionnaire for experiment B1.

A score can be computed for the answers on the SUS by adding up the recoded answers (using 0 to 4 from the worst to the best answer) and multiplying by 2.5, resulting in a final score between 0 and 100. For the search tool the final score was only 66.9, which would be below the mean score of 69.5 for SUS studies found by Bangor *et al* [16]. However, it should be pointed out that only a small number of users participated in the experiment and the nature of the evaluation was formative, thus the score should not be used to compare the search tool to a polished system. It can only serve as a baseline to use in follow-up experiments testing improved versions of the search tool. Instead the questionnaires were primarily used as a way to drive the investigation and aid in analysing the recordings for particular problems (e.g., if a participant expressed negative views on a question this was further pursued in the final interview with that user and while watching the video particular care was taken to find the reasons for this users problems with the UI).

The analysis of the second questionnaire reveals particular problems for questions 3, 7, 8, 10 and 14, with question 7 (*Program informs about its progress*) the one with the most negative answers. The system would benefit from additional feedback on its progress as well as better feedback in error cases (question 8) and more consistent and clearer labels. Excellent marks were given for system speed and system response time (question 11). This answer reflects the best case scenario in which the search itself (which is not part of the WP3 evaluation) is not considered. For the evaluation, results were cached and mostly instantaneous.

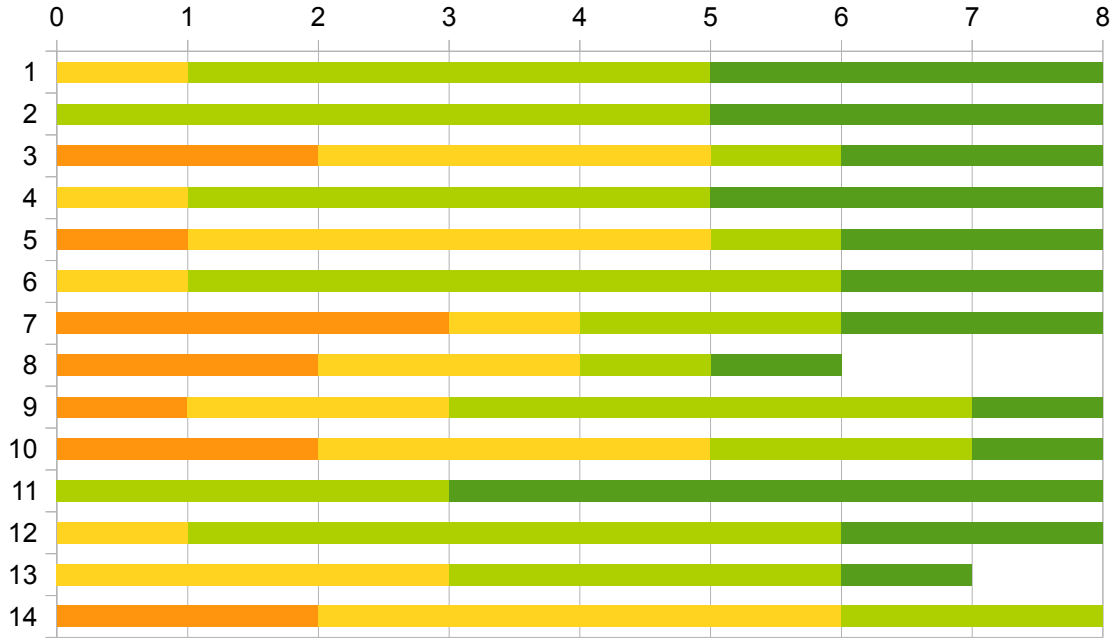


Figure 10: Results of the QUIS based usability questionnaire for experiment B1.

3.2.2 Description of Search Tasks

For experiments B2 and B3, four search tasks based on real user needs were created in a joint effort of tasks T3.6 ("Evaluation of the User Interface and Search Specification System") and T7.2 ("User-centered Evaluation Strategy"). For these search tasks, relevance criteria were defined and between 40 and 43 result documents (web-pages) were assessed (some were discarded as unsuitable for inclusion or as duplicates). Care was taken to select search tasks and results that covered multiple aspects, not all of which were necessarily relevant to the task. The results were coded for ezDL and fixed result lists were created from them. The four tasks selected are as follows:

Aspirin for children (Reye syndrome)

Description of task: "Your 9 year old kid has a head ache. You want to give him medication, but you've heard that aspirin is not good for children. You now want to confirm this and maybe find out why and what alternatives you have. Documents that are just about aspirin poisoning in general or overdosing are *not relevant*. Documents that are about pregnancy and aspirin are also *not relevant*."

Query terms: aspirin kids

Total number of results used: 40 (16 relevant)

Car accident related to sleep apnoea

Description of task: "You heard of a car accident in China, where a physician said on TV that it was caused by someone having 'apnea'. You want to know what this disease is and how it could have caused the accident. Treatment options are not of interest to you and therefore *not relevant*."

Query terms: accident apnea

Total number of results used: 43 (21 relevant)

Angelman syndrome

Description of task: "Your paediatrician told you that your daughter of 12 months has 'Angelman syndrome'. You want to know more about how that will effect her life span, quality of living and general prospects for the future. What are tested treatment options that are approved by the medical community? As you do not feel qualified to read research articles on the topic, those are *not relevant*."

Query terms: angelman syndrome

Total number of results used: 40 (16 relevant)

Screening for prostate cancer related to age

Description of task: "Your family of European descent has a history of cancer. You want to know more about correlations of cancer development and age, screening options and when to start screening, in particular for prostate cancer. You are not interesting in articles that merely describe the screening process, are about cancer screening without any reference to the age of the patient, scientific research articles or guidelines for health professionals. Those are *not relevant*."

Query terms: starting age screening cancer prostate

Total number of results used: 42 (29 relevant)

3.2.3 Experiment B2: Eye-tracking of Result Item Representations

Using the tasks described in Section 3.2.2, a second experiment was conducted that would examine the effect of different visualisations for result surrogates. It was also intended to find out which of them users would prefer. In the definition of use case requirements summarized in deliverable D8.2 a number of features were listed as being necessary parts of a link description. The importance for a good link description was stressed as one of the findings from the questionnaires. The aspects differ between use cases, but a complete list would contain:

a) type of content,
b) title or page name,
c) snippet
d) author,
e) publisher
f) URL,
g) target audience,
h) quality accreditation marks (if available),
i) quality rating and/or social ratings,
j) time of last update,
k) indication of free or restricted access,
l) and type of media.

Since there is a good chance that a result surrogate containing the complete list of information would be of overwhelming information density to most users, the experiment was designed to find out which

of the listed aspects are most often looked at during a search task. For this, a total of four result surrogate variations was designed (a picture of the first mock-ups can be found in Figure 11). All of the variations share a common area on the left side that shows: the rank of the document within the result, a visual representation of the calculated relevance (a green bar filled to a percentage that corresponds to the normalized RSV), and an icon that shows the media type. In addition, a common query term highlighting would be used for terms appearing in the title or the snippet of a result document.

1. The **first variant** is the baseline representation used in ezDL. It has a thumbnail for the result on the left if such is available, the document title on the first line, author information on the second, publication date and source on the third, and finally a short snippet from the document.
2. The **second variant** is modelled after the common result representation in web search engines. It is expected that users will find it familiar and easy to use. This variant does not show a thumbnail. The first line shows the title of the document in blue colour, below that a short snippet from the result document – including the date of publication or last update. On the line below the snippet the authors, or if no author are available, the URL of the document is shown. Finally (in an extension of web search results), a number of labels are presented (these are categories used within the Khresmoi project that were assigned to the documents or the websites on which the documents can be found). Information about the target audience is included among those labels.
3. The **third variant** has the highest information density and shows several descriptive aspects on each line. Again, the first line contains the title, the second line the date of publication or update, as well as publisher and author information. The third line presents a quality rating as stars (if available), information on access restrictions and the target audience. Below that are a snippet and a last line with other informative labels about the document.
4. The **fourth variant** is a combination of different features, using colour to offset the author line and including several thumbnails for images in the result document. Like the third variant it shows ratings if available. Labels are given on the same line, and also include labels that describe the target audience.

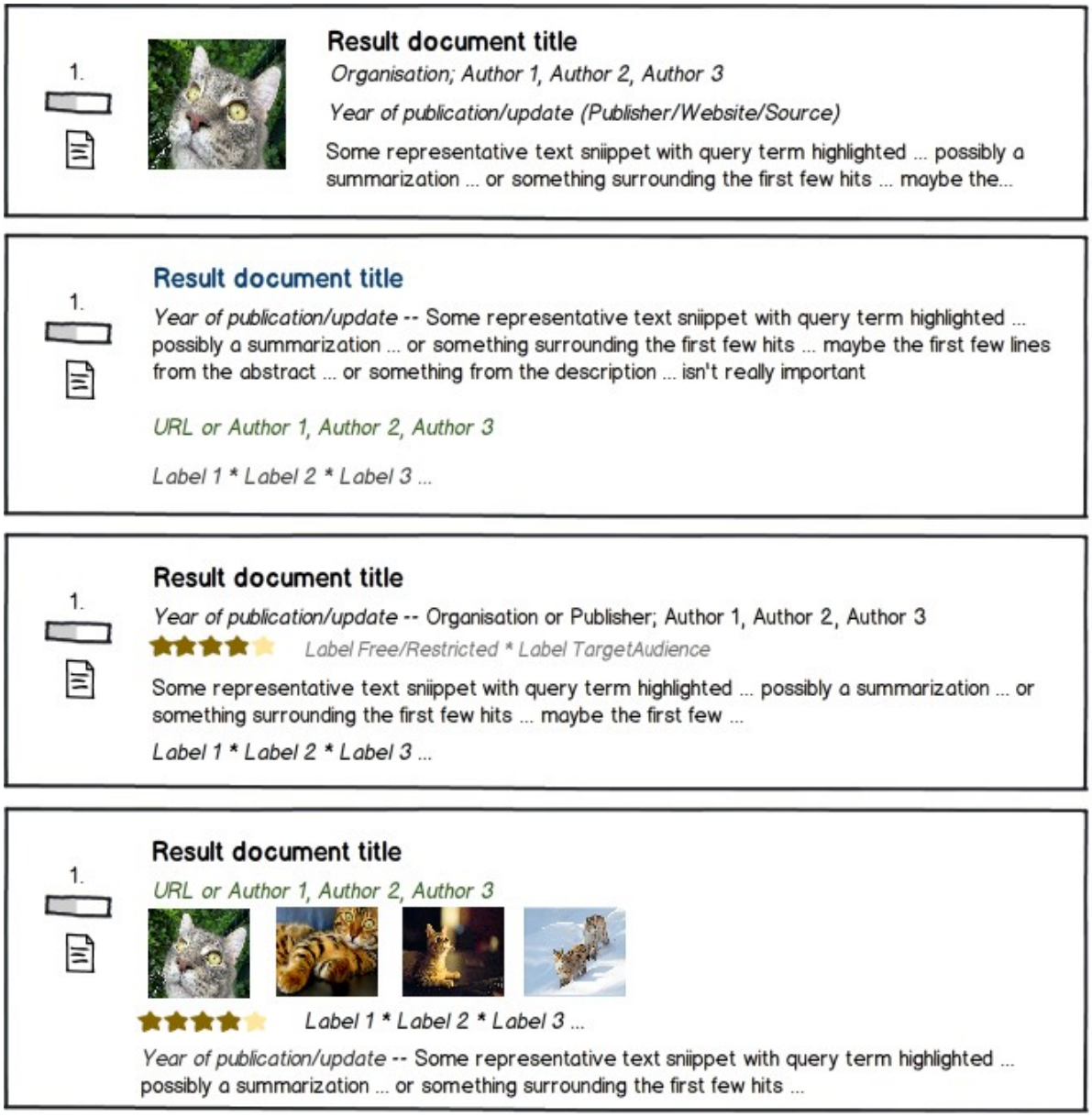


Figure 11: Mock-up of different result item presentations.

In the experiment users were asked to use these different result surrogates to decide for a given result list which of the presented documents are relevant to the given search task. The experiment was supposed to answer a number of research questions:

- Is there a significant difference in errors that users make between the different presentation variants? Do users miss more relevant documents, or judge more documents as falsely relevant using one of these variants than using the others?
- Which parts of the result representation are looked at most often or longest (relative to its size) during examining a search result? This would be determined by using an eye tracking system to assign Areas of Interest (AOIs) around each part, and counting fixations.
- Which parts of the result representation are found to be most useful by the searchers doing the tasks?
- Is there a clear preference for one result representation, or is the best result representation

dependent on the search task?

In preparation for the eye tracking, pre-tests were conducted to find the best resolution. An eye tracker can only determine the actual point of fixation with a certain accuracy. The typical error is about ±0.5 to 1° visual angle. This translates to about ±1cm for a viewing distance of 50 centimetres and makes examining adjacent or very thin AOIs difficult. The RED eye tracking system that is used in the usability lab at the University of Duisburg-Essen uses a Dell 20" screen with a default resolution of 1680x1050. With the default resolution the measuring error translates to over 40 pixels around the measured fixation point.

Since the result representations examined contain several small, adjacent areas a lower resolution had to be found. As a result of the pre-tests, a resolution of 1152x864 was settled on as a compromise between pixel accuracy and a realistic test setup. For practical reasons and to allow an aggregated evaluation of the gaze tracks and fixations, static results instead of dynamic result lists were used. This also eliminates any variations introduced by interaction of the participant with the list (e.g., different ways to scroll through a list).

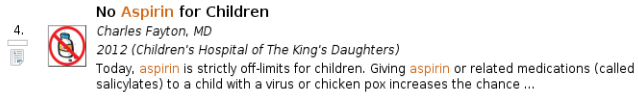
Screen shots of result list parts for each of the four search tasks combined with each of the four variants were created at the resolution of 1152x864. Each result list part showed between 4 and 7 results depending on the vertical size of the result presentation. For each task/variant combination a static task page was prepared that contained the task description on the left, together with the query keywords used, and the screen shot of the result list on the right. A Latin square design was used to combine the 16 combinations into a testing sequence.

A total of 16 participants were tested, aged between 19 and 32 (mean 25.5). Half of the participants were female, most (12) were students at the University of Duisburg-Essen. All were native or near native speakers of German and were presented the task descriptions in German. Thirteen had good, the rest adequate English skills. All were able to understand the short texts (title and snippets) in English. Labels and indicators for target audience were translated to German. Of the participants, 4 did not use web search on a daily basis (3 of these only rarely search the web). All but two had previously searched for health related information on the web, but only two did so often. All participants were familiar with Google, ten had used other web search engines (Bing).

During the tasks the participants were encouraged to examine the result lists in a normal speed. If they found one or more results in the list relevant to the given search task, they stated the numbers of the document. This was noted down by the experiment facilitator. The participants were able to advance the experiment themselves by looking at a button on the screen. This would bring up the next of the 16 tasks.

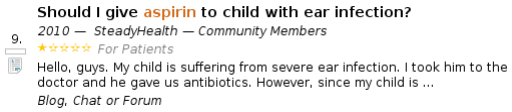
After the experiments the participants were asked to fill out a questionnaire (presented in Appendix 7.5). They were also given a print out with examples of the four variants (see screen shots below) to refer to while they were filling out the questionnaire. The different parts of the result representations were marked with numbers and the same numbering scheme was used in the questionnaire:

- **Variant A:**



- **Variant B:**

- **Variant C:**



- **Variant D:**

When asked for their personal preference on the different result presentations two clear favourites emerged: the **variant B**, which was modelled after the familiar web search results, was named by 6 as a favourite, and the information rich **variant C** was named by 9. No one preferred the baseline representation, one participant the variant with thumbnails (she liked that this variant showed the URL of the result and was neutral on the thumbnails).

Result aspect	avg. fixation time	avg. number of fixations	percentage of all fixations	avg. number of re-visits
Variant A (baseline, existing ezDL presentation)				
Text snippet	3.25s	10.30	26%	5.13
Document title	1.39s	4.70	12%	2.82
Publishing organisation/authors	1.46s	4.85	12%	2.75
Thumbnail	.27s	.97	1%	.45
Publication date	.32s	.98	2%	.38
Variant B (web search based result presentation)				
Text snippet	4.05s	12.71	27%	6.12
Document title	2.26s	7.43	16%	4.44
Publishing organisation/authors	1.03s	3.47	8%	2.11
Target audience and other labels	.73s	2.46	5%	1.35
Publication date	.43s	1.35	3%	.68
Variant C (information rich presentation)				
Text snippet	3.16s	9.77	29%	4.08
Document title	1.29s	4.47	13%	2.35
Publishing organisation/authors	.63s	2.33	7%	1.29
Target audience and other labels	.45s	1.54	7%	.79
Publication date	.34s	1.11	3%	.5
Variant D (presentation with multiple thumbnails)				
Text snippet	3.39s	10.39	25%	4.20
Document title	1.33s	4.34	10%	2.42
URL	1.09s	2.64	9%	1.97
Target audience and other labels	.64s	2.11	5%	1.11
Thumbnail	.60s	1.85	3%	1.02

Table 1: Eye-tracking data for result aspects (not all documents for A and D had thumbnails).

Next we looked at the answers about the usefulness of specific aspects of the representations, which are summarized in Table 2. Since it is most interesting whether a specific aspect is useful to include in a result item representation or not, the two positive values for each Likert item were conflated and are represented by a percentage (i.e., for "text snippets" 86% found them helpful or very helpful). The resulting percentages were then arranged in descending order, resulting in a ranked list of result aspects. Clearly the most important parts for the searchers were **snippets**, **highlighting of query terms** and the **document title (Figure 23)**, which is not very surprising. Of more interest are the importance of **target audience**, **publishing organisation** and to a lesser degree the **URL**. Here the search domain clearly plays a big role.

Of little to no importance are the relevance bar, the rank number of the document (probably an artefact of the artificial experiment setup, since normally users tend to trust in the ranking of search engines and assign high importance to the document rank) and thumbnails.

For three aspects of the result presentations participants mentioned that they did not know what kind of information was shown: relevance bars were mentioned by 4 participants, document rank and thumbnails by 1 each. Relevance bar are more typical of digital libraries and professional information retrieval systems than of the web search engines familiar to most of the participants. Showing them does not offer any real benefit to the searcher and they will therefore be removed in future versions. The confusion regarding the thumbnails probably stems from the perceived unrelatedness of some of the shown images to the search task. The thumbnails are taken directly from the result page, which can sometimes be about multiple topics or contain images that are previews for navigation links to other parts of the site.

Result aspect	Is helpful for judging relevance...
Text snippet	**87.5%**
Highlighting of query terms	**87.5%**
Document title	**81.2%**
Target audience	**62.5%**
Publishing organisation	**50%**
URL/Link	37.5%
Authors	18.7%
Publication date	18.7%
Other labels	18.7%
User ratings	12.5%
Rank	6.2%
Thumbnails	6.2%
Relevance bar	0%

Table 2: Helpful result aspects.

Not all eye-tracking data collected was usable. Some participants' eye movements could not be reliably tracked. Using the eye-tracking data of the 8 usable recordings, AOIs were defined around each element of the result items (Figure 25). The data was then analysed for: total time of fixation, number of fixations and number of revisits per AOI group (all snippet were aggregated in one group, all titles, all authors, etc.). The results of this analysis are summarized in Table 1.

Unfortunately, even with a low resolution reliably capturing the exact areas of fixations is difficult if the areas are small and close together, as is the case for the individual parts of a result representation. This needs to be kept in mind, when looking at the collected data. As it is, the objective eye tracking data supports the subjective statements of the participants about which result parts where most important for judging relevance: text snippets is clearly the most important and most often consulted part. More indicative than the time spend reading is the number of revisits. While longer texts clearly also take more time to read, revisits indicate that searchers came back to this element to check. It can also be seen (Figure 24) that thumbnails and publication date are only rarely consulted and almost never revisited – again confirming the answers given in the questionnaire.

Few differences can be seen between the variants for most included result parts. The thumbnails received more attention in variant D compared to variant A, which can be attributed to the larger number of thumbnails included in the first variant. The only difference that sticks out is the large percentage of fixations for author/publishing organisation in variant A. This was the ezDL baseline that didn't include information like target audience, other labels or result URL. Here the searchers seem to have focused more on the little information available and may have tried to glean e.g. the target audience from the publishing organisation (this is somewhat corroborated by participants' comments made during the experiment).

Finally, the number of errors was tabulated for each combination of task and result list variant. Errors can be of two types: false positives (FP), e.g. documents that a participant deemed relevant but which were not, and false negatives (FN), e.g. documents that were relevant but were ignored by the participant. While false negatives lead to searchers potentially missing information that they were looking for, false positives waste the searcher's time and can lead to frustration with the search engine. Table 3 summarizes the rate of false positives, rate of false negatives and total rate of errors for all four result variants.

An Analysis of Variance (ANOVA) was used to detect correlations between result variant and error rates. A correlation was found between result variant and rate of FP and between result variant and rate of total errors (both with $p < 0.01$). To find out which variants performed significantly better than others, an ad-hoc pairwise comparison of the error rates for each result variant was performed using a paired, two-sided Wilcoxon rank-sum test (aka. Mann-Whitney U). The results are summarized in the Table 4 with * denoting results that were significant at *.05* and ** denoting results that were significant at *.01*.

The average number of FP errors did not differ much between variants A (17.36%), B (19.79%) and C (16.25%). For variant D the rate of false positives was significantly higher with 35.42% ($p < .05$ for all comparisons), which may have to do with the high number of thumbnails. Thumbnails are visual distractions which in this result variant also divide one result in two parts (title and author/URL above the thumbnails, rest below). The images from the result page may move attention away from better indicators of relevance and lead searchers to believe that the page is more relevant than the textual snippets would have.

Rate of...	FP	FN	FP+FN
Variant A	17.36%	53.52%	40.50%
Variant B	19.79%	46.15%	33.50%
Variant C	16.25%	51.88%	30.50%
Variant D	35.42%	43.01%	41.88%

Table 3: Error rates per result variant.

On the other hand, the average number of false negative errors was lowest for variant D, which again might indicate that the searchers tended to judge more results as relevant when they were presented with thumbnails. However, the error rate for false negatives had a high variance and the ANOVA found no clear correlation between the result variant and the rate of FN errors.

For the combined number of errors the data was far clearer. The ANOVA found a correlation between result variant and error rate with $p=.001$. Pairwise comparison of the values for the four variants showed that **variant B** is significantly better for reducing the total number of relevance judgement errors than variant D, and that **variant C** is significantly better than A and D (there were no significant differences between B and C or A and D).

FP FP + FN	B	C	D
Variant A	2.43 / .253 7.0 / .101	1.11 / .753 *10.0 / .018* *	*18.06 / .009* ** 1.38 / .649
Variant B	-	3.54 / .154 3.0 / .378	*15.63 / .018* * *8.38 / .010* *
Variant C	-	-	*19.17 / .006* ** *11.38 / .006* **

Table 4: Pairwise comparison using two-sided Wilcoxon signed rank test (difference and p-Value).

These results confirm that variants B and C, which were also favoured by the participants, are best suited for judging the relevance of result documents among those variants tested. Of those two, variant C seems to offer a slight advantage, as it was favoured by more participants (9 vs. 6) and performed significantly better than both variants A and D (while B only showed a significantly better performance than D). Going forward, variant C will be used as the main representation of result items, incorporating some of the features used in B such as different colours for different parts of the result item which will make it easier to quickly find the parts at a glance. Based on the favoured variant C, additional variations could be designed and tested in the following months or a configurable result presentation could be offered.

3.2.4 Experiment B3: Presenting Results in Grouped Lists vs. Tabs

In the next experiment we wanted to find out which of two different result presentations was preferred by the users. Both visualize categories within the result set, one using separate tabs for each category and one a grouped list combined with a navigation to the left. Figure 12 shows the result presentation with tabs. There is one tab for each document label and because of space constraints this variant only shows the most important categories as well as one tab with the remaining documents. Figure 13 shows the result variation using a grouped list, where each group shows documents for a particular label. The header of a group shows its name as well as the number of documents in this group. Both methods allow for documents to appear in multiple categories, e.g. a document that is in categories “for health professionals”, “women's health”, and “guidelines”.

Each participant worked on the same four tasks used in previous experiment (see the task description in section 3.2.2). For each task either the grouped variant or the tabbed variant was shown. The mapping from task to result presentation variant was rotated according to a standard Latin square design, so that each combination appeared the same number of times, and each variant and each task appeared in each position during the experiment. The users were instructed to collect relevant document in the tray tool. In contrast to the previous experiment the complete and interactive result lists for each task were presented to the user. All result lists used the same variant to display individual result items (the baseline variant A, see Section 3.2.3). The following categories were present within the result sets (the underlined categories were used for tabs, while all categories were presented in the grouped list):

- **Labels for task 1:** “general population/for patients”, “signs and symptoms”, “prevention”, “treatment”, “drugs”, “side effects”, “hospital/clinic/medical centre/medical establishment”, “chat/blog/forum”, “dictionary definitions”, “kids”, “adults”, “tests”, “images”, “women's health”
- **Labels for task 2:** “general population/for patients”, “health professionals”, “research”, “dictionary definitions”, “hospital/clinic/medical centre/medical establishment”, “genetics”,

"treatment", "tests", "signs and symptoms", "differential diagnosis", "causes and risk factors", "scientific publication/research paper", "clinical trials ongoing", "online book", "faq"

- **Labels for task 3:** "general population/for patients", "warning and recalls", "dictionary definitions", "treatment", "tests", "signs and symptoms", "forum/message board", "signs and symptoms", "causes and risk factors", "hospital/clinic/medical centre/medical establishment", "clinical trials ongoing", "videos", "women's health", "men's health", "continuing education/cme", "kids", "links", "publication of clinical trials", "news", "seniors", "research", "scientific publication/research paper"
- **Labels for task 4:** "health professionals", "general population/for patients", "hospital/clinic/medical centre/medical establishment", "tests", "news", "patient support groups", "women's health", "men's health", "practice guidelines", "procedures", "causes and risk factors", "research", "scientific publication/research paper"

A total of 12 participants were tested, aged between 19 and 34 (mean 26.5). Of the 12 participants 5 were female, most (9) were students at the University of Duisburg-Essen. All were native or near native speakers of German and were presented the task descriptions in German. Ten had good, the rest adequate English skills. All were able to understand the short texts (title and snippets) in English. Labels and indicators for target audience were translated to German. Of the participants, all but four used web search on a daily basis. All but two had previously searched for health related information on the web, but only one did so often. All participants were familiar with Google.

During the tasks the participants were encouraged to examine the result lists in a normal speed. If they found one or more results in the list relevant to the given search task, they could use drag & drop or the context menu to save the document to the tray tool included with the system. The participants were able to advance the experiment themselves by clicking on a button. This would end the task and clear the clipboard. The next button click brought up a new task description, and another the result list. All documents saved to tray, as well as all interactions with the result lists (scrolling, switching categories or tabs, viewing details) were logged using the logging framework described in Section 2.1. After completing all four tasks the participants were given a very short questionnaire in which they could grade the two list variations and state a preference.

Of the 12 participants, 11 stated that the grouped list was helpful (5) or very helpful (6) for finding exactly those results they wanted for their search tasks. For the separate result lists this was stated by 6 participants (4 helpful, 2 very helpful). When asked to name their preferred method for presenting the different facets of the result list, 9 participants named the grouped list, 2 the tabbed list and one participant did not want to commit to either.

During the experiment each participant was presented with all results for the four search tasks described in Section 3.2.2, a total of 165 results of which 82 were relevant and should have been saved by the searchers. The presentation which showed either only parts of the result list, or a single result list with several grouped sections, made it most interesting if participants missed more result with one of the two variations. Over the course of the experiment both list variations were used to present an equal number of relevant documents (492). Using the grouped list, participants missed 59.2% of all relevant documents (on average 24.3 per participant), while using separate tabs users missed 65.6% (26.9 documents) of the relevant documents presented to them. However the difference is not significant using a Welch two sample t-test ($p=.27$).

Further evaluations using the collected data will need to be done, e.g. comparing time needed for the relevance judgements using both list variations and analysing the logs to see if and how the lists were used to skip irrelevant sections or facets of the result.

Figure 12: Result presentation variant with tabs.

Figure 13: Result presentation variant with grouped list.

3.3 Evaluation of Collaborative Components

As one of the results of the use case analysis regarding the general public as well as physicians, it became clear that users wished for collaborative features that were not initially envisioned to be part of the interface. The extension of the user interface by a collaborative aspect was also pointed out as a recommendation during the first year review of the project. Collaborative features or functions identified as requirements in deliverable D8.2 were:

- the possibility to rate documents according to their quality
- ranking results by social ratings and seeing social ratings within results
- query completion based on query history
- storing documents in a self-made compendium or keeping a list of favourites
- sharing results with physicians or peers
- having a forum for discussion or possibly a way to chat with other users

In addition, the possibility for tagging or annotating stored documents, as well as discussing or chatting about shared documents was raised as a valuable addition to the system. Since most of these features went beyond the initial DOW, first work on them started in early 2012. Here we present the extended plans for an interface that supports collaborative work on health documents found through the system and initial findings of a usability evaluation of some of those features.

3.3.1 Description of Planned Collaborative Components

Based on the early findings about user needs within the Khresmoi project (i.e. deliverables 8.1.1 [10], 8.1.2 [90], 8.2 [12], and 9.1 [13]), a number of tools were identified that could enhance a medical search engine with collaborative functionalities:

- Result rating or quality ratings of documents, and display of ratings (from D8.1.1: "*54% think it would be useful to directly rate search results and view other users' ratings.*", p. 27; from D8.1.2: *"When asked about which tools they preferred an interesting result was that the most important tool was 'being able to quality rate information/websites and perceiving ratings of other physicians'"*, p. 34)
- Sharing of knowledge (from D8.1.2: "*specialised, secured physician communities where we can exchange knowledge about patient cases with other physicians*", p. 36; from D9.1: *"[...] the communication among colleagues is used to share this knowledge not only during training, but also in clinical practice. Past cases store experience of other colleagues and could make this experience available in a more systematic way."*, p. 20)
- Keeping a history of previous searches (from D8.2: "*Khresmoi tracks the search history.*", p. 13 and other places)
- Keeping a personal hot-list, a list of favourites, or a personal collection/compendium (from D8.2: "*A possibility to save search results into a repository for all queries, audio and video files.*", p. 35)
- Sharing documents or results with physicians or peers (from D8.2: "*Possibility to share the results with a physician/friend or peer*", p. 35)

Goeuriot *et al* [8] describe the plans to support collaboration, especially for improving resources in the Khresmoi system (covered in the upcoming deliverable D3.4). To allow for providing these social functionalities, a number of additional components become necessary parts of the user interface:

- a personal, customizable user profile;
- a function to search for other users and possibly "friend" them or keep a roster of known users;
- and creating and managing groups of users.

The available categories of users (e.g. member of general public, medical student, hospital staff member, independent general practitioner) will be configurable by the system administrator of the

Khresmoi search system, who will also be able to set policies on who will be able to annotate, rate or collaboratively edit documents, or which categories will be able to communicate and share using the system and in what fashion.

By using ezDL as a base for the Khresmoi user interface, a history of previous searches was already provided. It was decided that initially a personal facility for storing and tagging interesting documents, and a way to share those documents with other users would be added to interface as part of task T3.2 "Query specification support, result presentation and personalization" (see Figure 14 for an early-stage mock-up). In addition, user profiles, user search and group management components would be provided as part of task T3.1 "Flexible user interface framework" (see the user profile mock-up in Figure 15). Their usability would be evaluated during the first component evaluation phase.

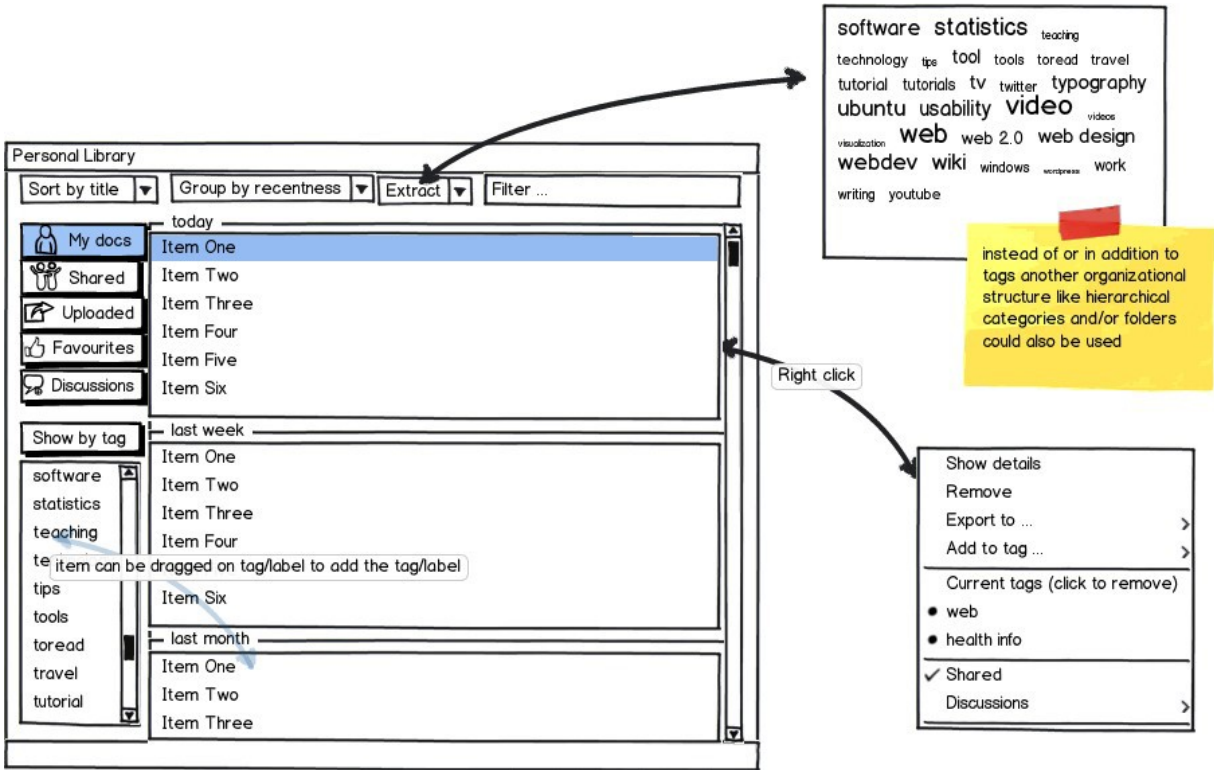


Figure 14: Initial mock-up of a personal document store.

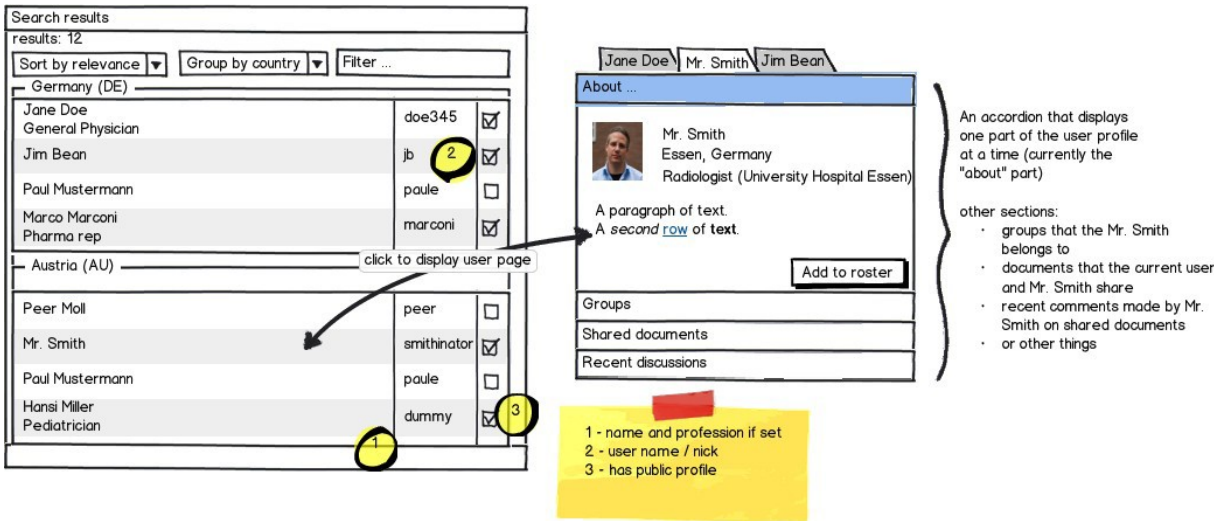


Figure 15: Initial mock-up of user search and profile.

3.3.2 Experiment C: Usability Test of Personal Collection Management and Sharing Facilities

During the month of July 2012, a first usability evaluation of the personal collection management and sharing facilities was conducted. This evaluation took place at the University of Duisburg-Essen and at Dublin City University. Experiments at the University of Duisburg-Essen were recorded with eye-tracking equipment (screen-recordings, mouse-movements, and gaze data). Complete logging data of the user interactions was collected for all experiments. In addition, one or two facilitators kept notes during the experiments and performed a short interview with the participants.

During the experiments, a total of 9 participants were asked to perform a number of typical tasks with the user interface that covered saving favourite documents to a personal collection and deleting documents, tagging results and removing tags, sharing results with other users, creating a personal roster for sharing, searching for users, re-finding saved documents by filtering and grouping within the collection, and managing the personal user profile. The participants were given no introduction to the user interface and did not receive help during their tasks (unless they were stuck and could not continue). All participants were initially unfamiliar with the ezDL or Khresmoi interface. The goal of the experiments was to discover usability and learn-ability problems that need to be addressed before continuing development on the collaborative functions. Since the social features are advanced functionalities, it is to be expected that searchers who are unfamiliar with the systems will initially need help e.g. in form of an interactive tutorial or an introductory video or animation. Thus, an additional goal of the evaluation was finding those interaction areas that would need such support.

A complete description of the tasks for this evaluation is given in appendix 7.3.

Three of the participants were female and 6 were students at various stages of their studies (graduate and undergraduate students). 5 used the German version of the interface, with the remaining participants using the English version. Before the actual tasks, the participants were asked about their previous experience with computers and web search in general, but also about their experience with systems that provide similar functionalities as those under consideration.

For the aspect of storing and managing documents, desktop file manager (Windows Explorer, Mac

Finder, Nautilus, and similar programmes) as well as online services like Mendeley[1], Bibsonomy[2] or digital libraries with personal portfolio were considered. Six participants used file managers on a regular basis, two sometimes and one never. Only one participant was familiar with online systems that allow storing and management of documents.

For the social aspects the participants were asked about their experience with various social network sites (Facebook[3], Google+[4] or similar sites). All of these allow management of a personal profile, while some allow for sharing of documents and some for management of personal contact groups (e.g. Circles in Google+). All nine participants were regular users of Facebook, all of them sometimes or often used other specialised social networks (e.g. StudiVZ[5] which caters specifically to German students) and 7 had used Google+.

Only two participants were able to successfully complete all the tasks without help. On average, users needed 16 minutes and 9 seconds to complete the experiment (between 9 minutes 53 and 25 minutes 50). Half of the users had to abandon *tasks 8-10* ("search for a named user, view his profile and add him to your contact group") since they were unable to find the user with their search.

The longest task on average was *task 4* ("re-find the documents you just saved and add a tag"), which took the users an average 150 seconds to complete. The most time-consuming problem for two users was finding the personal collection, even though it was open from the beginning and covered a third of the screen real estate. One reason for this might have been a label mismatch, since the users were not able to recognize the open tool called "personal library" as the place where they had previously stored their documents. Another reason given by one user was that she had not expected the personal collection to already contain documents, and that she had therefore ignored it and searched for some other tool that only contained those two documents she just saved. Re-finding the documents was almost always accomplished by applying a filter to the personal collection, or by scrolling through the list of documents. Only one participant sorted the list by "date of addition". Tagging the documents itself was usually unproblematic.

It is interesting to note, that while the users did not sort by date of addition to find the newly added documents, they did group by date to find documents added "in the last month" for *task 11*. It seems likely that the position of the sorting menu (which was different from the position of the same element in the search tool) caused many of them to miss it, while the grouping menu was easily recognized.

Creating a private contact group for sharing (*task 6*) and adding two users to it (*task 7*) uncovered a large number of small usability problems, and some display bugs. Despite this, all users were able to complete these tasks without major difficulties (the fastest user needed 25 seconds to find the group creation and create a new group, the slowest 112 seconds; adding the two users to the contact group took between 37 and 109 seconds). In addition, five out of six participants were able to share a document with two other users without help (*task 5*).

As with experiment B1 a problem protocol was created by the experiment facilitator and observer. After the experiments five developers reviewed the video material and expanded the problem protocol from their observations. The usability problems that were discovered are currently being addressed within tasks T3.1 and T3.2:

- Six users had problems finding the settings for the personal profile. The placement within the file menu was unexpected and the label for the menu entry ("options") did not clearly indicate

[1] http://www.mendeley.com

[2] http://www.bibsonomy.org

[3] http://www.facebook.com

[4] http://plus.google.com

[5] http://www.studivz.net

that it could be used to change profile settings. Suggestions were to have a menu option clearly labelled as profile or to use the web convention of having the user name as a link to the profile [Fixed as of Aug 2nd.]

- Five users had problems finding the user search. The after-session interview revealed doubts that the same query prompt that was used for searching documents could also be used for searching users. To improve this, a toggle was added that switches between searching everything and searching only for user profiles. The state is clearly indicated. [Fixed as of Aug 30.]
- Three users scanned the menu and the status bar for the personal library although it was already opened and occupying about one third to the screen.
- The difference between saving to the temporary tray/clipboard and saving to the personal library is not intuitively clear.
- The auto-completion feature for entering user names or groups was not used as often as expected as participants tend to not look at the screen while typing.
- The group tool lacks a context menu which half of the users expected it to have. Also, one participant raised the suggestion of combining the group list and the group edit view into one view.
- Adding other users to a group requires knowledge of their login names instead of their real names.
- One participant tried to drag and drop a user detail page into a group. This functionality might be added as an additional means for achieving this task.
- It seemed to irritate participants that the group tool requires them to explicitly select a group to refresh its member list although the group has been selected before and is still highlighted. [Fixed as of Aug 8th]
- Participants expected to be able to perform a user search directly from within the user group tool. Some managed to use the normal search box to complete the task others needed advice.
- Finding the profile of a user was hard for some participants because it required them to search for the user and then click on the entry in the result list. Three users abandoned the task since they were unable to find the user within the search result.
- Participants often expected the filter field in the group tool to trigger a search for users. Even though the filter field looks just like other filter fields in the program, which were successfully used and recognized, apparently the position and the lack of affordance lead them to misidentify this UI element.
- One participant did not understand the concept of a personal collection/library at all and mistakenly tried to use it for searching.
- A few participants were confused after finding the personal library initially filled with some documents. One participant stated that ”not all search result are supposed to appear in the personal library”.
- One participant suggested offering multiple personal libraries per user (presumably to have more of a separation between different work contexts than tags can offer)
- Participants expected to be able to search for users in the “add to group” dialog.
- The experiments also revealed that the visual feedback for adding a document to the personal library was hardly ever seen. New documents are added at the position that is correct for the

current sort order. If the collection is already larger than what can be displayed this position might be below or above the current scroll window and therefore not visible. An easy solution is to automatically scroll to the position were the document is inserted.

- The drag and drop behaviour of the system should be unified across all tools.
- Almost all participants tagged or deleted documents individually instead of selecting multiple documents and performing the action on the complete selection.
- Participants wanted to be able to view the profile page for members in a contact group which currently is not possible.
- Four participants tried to print a document using the context menu before going to the detail view, where they found and used the print button. Even though this option is not present in many file managers it is apparently something users expect.
- When trying to add a contact to a group from the contact's profile, an empty selection is presented if the current user does not have any groups yet. It should be possible to create a new group from that dialog and directly add the contact to the newly created group.

After the experiments the participants were asked to fill out a usability questionnaire, the results of which are summarized in Figures 16 and 17. For the nature of the questionnaires and the methods used to code the responses in the bar charts see Section 3.2.1. Even though the collaborative components are at an early stage of development the feedback was not significantly more negative than the feedback received for experiment B1. Four of the participants answered that they would use such a system again, and 3 were neutral and 2 said they would not use such a system often. Unlike for the search tool (detailed in Section 3.2.1), participants remarked negatively on the integration of the various components. This complements the observations collected above, in that drag-and-drop behaviour, integrated user search or context menu actions were often expected but not yet present. Despite the existing bugs and problems, the users mostly found the system not difficult to use.

Figure 16: Results of the SUS questionnaire for experiment C.

Among the answers to the second questionnaire, those for questions 3, 4, 5 and 14 were the most negative. Half of the participants found the labels inconsistent (question 3) and two did not think that the terms used in the program were fitting to the task (question 4). Currently, new terms for some of the problematic features are being discussed. This is also something that has to be kept in mind for the translation of the interface (which is currently available in German, English, French, Spanish, Czech, Chinese and Vietnamese). Unfortunate translation can have a huge impact on user satisfaction if they obscure the meaning of tools and features, and make it harder for users to find what they are looking for. The very negative response to question 5 (which addresses the position of messages on the screen) can partially be explained by the setup of the eye-tracking experiment with two monitors (one for the participant, one for the observer). Due to a bug, some messages appeared on the observation monitor and had to be moved to the correct screen.

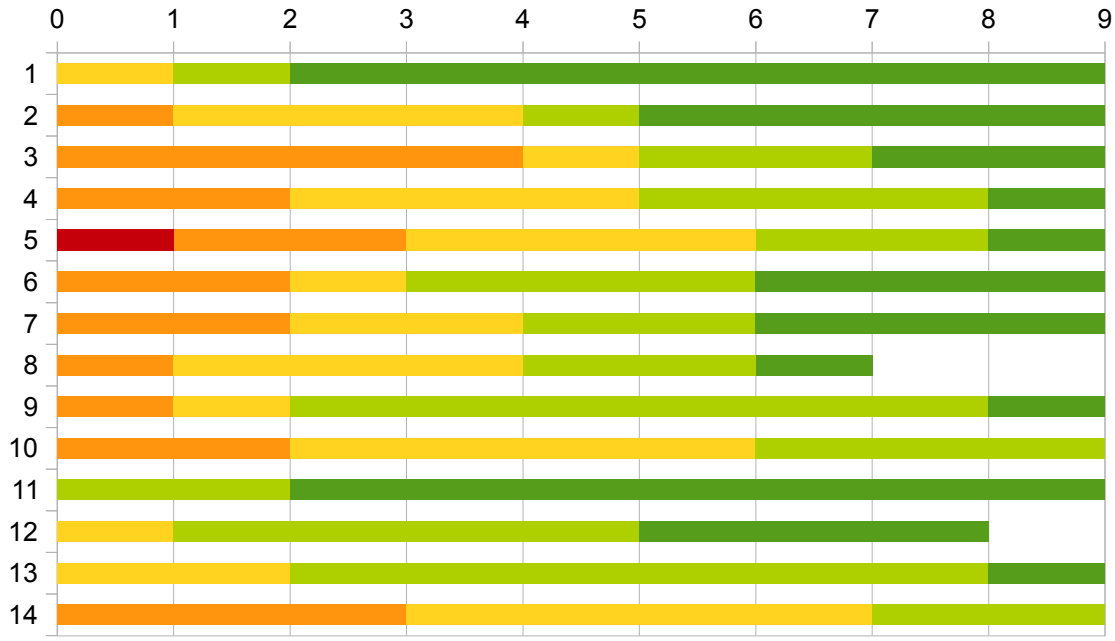


Figure 17: Results of the QUIS based usability questionnaire for experiment C.

From the interviews with the participants, combined with the results of the questionnaires it seems that there is indeed interest in collaborative and social elements in a health search engine and that the current features are similar to what users of the Facebook generation expect. While it is obvious that many usability enhancements still have to be done (the SUS score for the collaboration features amounted to 63.6, which according to Bangor *et al* [16] corresponds to a marginally acceptable grade), a social extension of the system seems worthwhile and not unduly confusing to those users who are most likely to use such features.

3.4 Evaluation of RIA Client

To access ezDL and therefore the Khresmoi system, a client is necessary. A number of different configurations and variants are available, most of them Swing-based Java desktop applications. Java

Swing allows building powerful and platform independent applications. On the other hand it introduces difficulties for users who are not very proficient with computers, as it requires them to have Java installed on their machines and to download and run an application. Many users prefer to abscond with desktop applications altogether and move more and more of their computer use into the web browser (see e.g. the growing suite of Google browser applications for e-mail, calenders, task management, document creation).

For those users a so-called Rich Internet Application (RIA) was developed as a second interface to ezDL[6] (see the screen shot in Figure 18). While there exists no formal definition of what an RIA is, such applications can be characterised by a number of main properties. RIAs use web technology to achieve a look and feel that is close to those of desktop applications but can be executed directly in a browser. They also offer interactive techniques like direct manipulation, drag and drop and context menus. The web client for ezDL was evaluated as part of a diploma thesis at the University of Duisburg-Essen [15].

The Khresmoi RIA user interface uses many techniques and interaction paradigms that were already evaluated in the previously described experiments: grouped result lists, filtering, saving documents, suggestions during query formulation (see Figure 19) and search for similar images. A strict component evaluation was therefore dropped in favour of a system evaluation of the first client prototype. The aspects of learn-ability and usability were examined in an experiment with 15 participants who were asked to perform seven tasks each. Five of the tasks were designed to find usability problems whereas two similar tasks were used to measure learn-ability (the tasks are described in Appendix 7.4). Time for task completion was also measured for all tasks. All participants filled out two questionnaires (reprinted in Appendix 7.5). The questionnaires contained Likert-scale items and were designed to cover all applicable criteria of the DIN EN 9241-110 standard for evaluation interactive systems. The following paragraph describes the findings of this experiment, originally published in [15].

The questionnaire items were aggregated into different categories to ease the analysis and allow for a clearer understanding of the results. For every category an optimal score was calculated. Actual results are presented as percentages of this optimal score and can be found in Table 5.

Category	Questionnaire 1	Questionnaire 2
Task suitability	88%	90%
Controllability	82%	85%
Positioning and visual appearance	85%	84%
Conformity to expectations	76%	78%
Error tolerance	84%	86%
Feedback	80%	77%

Table 5: Results for usability evaluation of RIA.

To find out whether the interface can be learned by first time users without any help, the execution times of two comparable tasks (1 and 7) were taken into account. The mean time to completion for task one was 348 seconds while task seven was in average completed in 228 seconds. This is an improvement of approximately 34%. A t-test yielded a statistical significance with a p-score of *0.002*.

Because of the small sample size these findings cannot be generalized. Nevertheless, they show a tendency that the system can be learned in a reasonable amount of time by users without prior

[6] A current version of the web interface is available at http://screwdriver.is.inf.uni-due.de:8182/gframedl-web-1.0.0-SNAPSHOT

knowledge of the application.

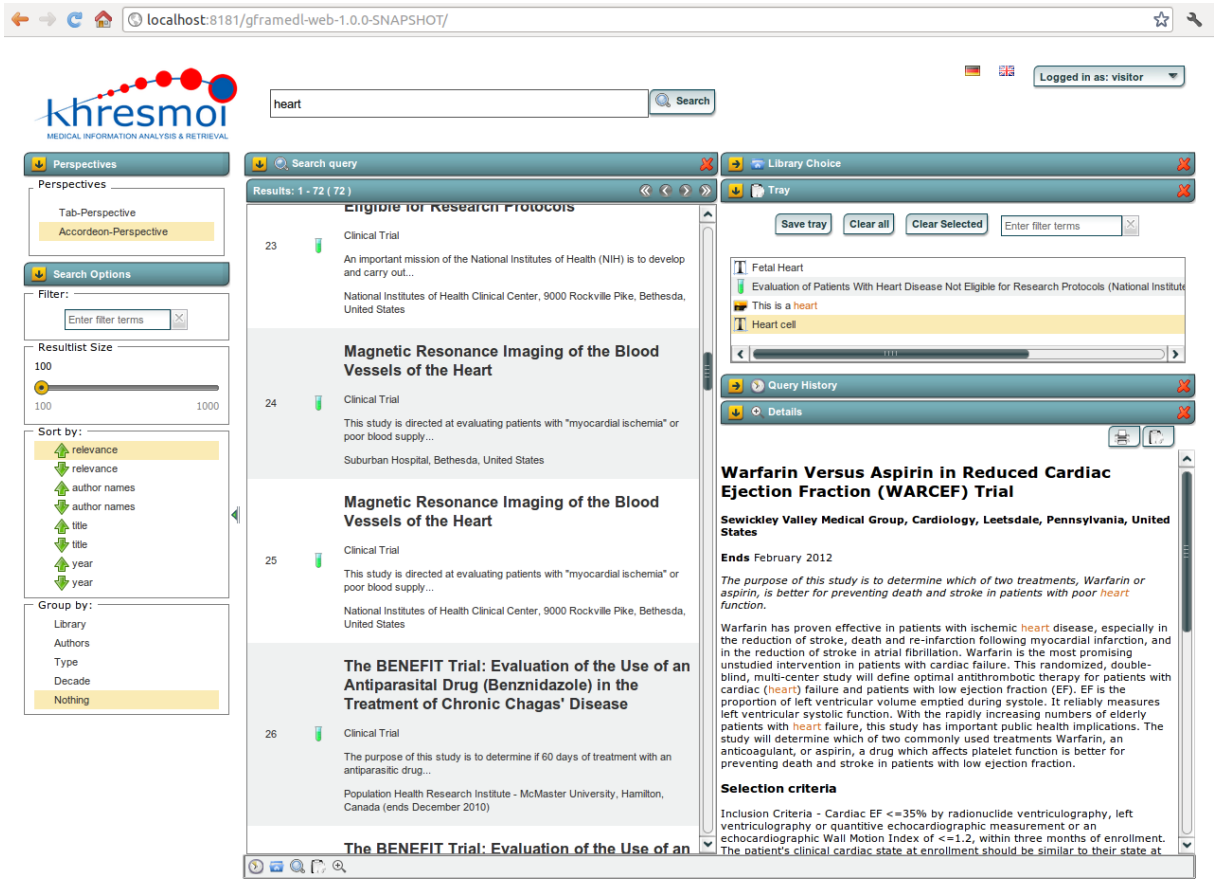


Figure 18: Screen shot of web interface.

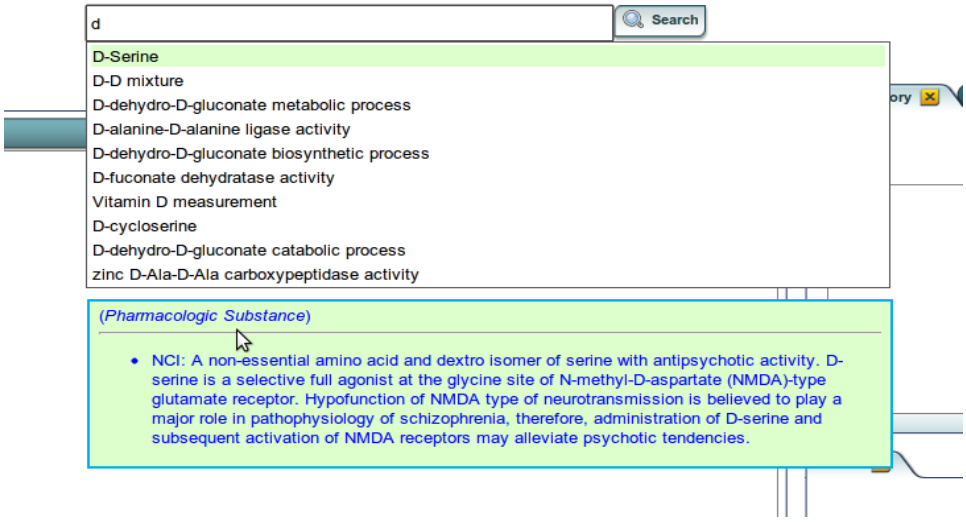


Figure 19: Screen shot of term explanations in web interface.

4 Discussion and Further Evaluations

Three usability experiments were conducted showing the general suitability of ezDL as the flexible user interface framework. Nevertheless several usability issues and inconsistencies have been discovered. Some of which already have been resolved. All remaining issues will be resolved before the WP 10 user centred evaluations are performed.

A number of experiments were performed for evaluation different user interface components such as result presentation, suggestions and collaborative tools. For result presentation participants made statistically significant fewer errors with two out of our four proposed variants. These variants were also preferred by the participants according to the questionnaires. The findings regarding presentation of result surrogates in a grouped list vs. tab-based list were not as valuable as the afore mentioned results. The evaluation of the collaborative components revealed usability problems but also showed general interest of participants in such features.

The evaluation of different designs for the presentation of search suggestions showed that none of the examined choices was clearly superior to another. This can be attributed to small sample sizes, but also to a possibly small effect size. A conclusion from this is that also neither design was clearly inferior, so offering both designs and leaving it up to the user might not have negative consequences.

The rich internet application user interface which uses the same general concepts as the full featured client was evaluated regarding learn-ability aspects. The experiment showed that the interface can be understood and used by participants without prior knowledge after short time. The evaluation also revealed fairly high usability scores of this interface.

The evaluations were conducted with an early prototypes of the user interface. These interfaces were neither feature complete nor free of issues and future development will address these problems. The usability test did however show parts of the full featured client where users would benefit from tutorials, something that will help in conducting the user-centred evaluations of the complete search system. Where the basic search functionalities were all easy to use and caused no problems for the users, it would be beneficial to provide introductions to advanced and unusual features like grouping results, extraction of terms, using the personal library, managing contact groups and sharing documents with other users.

New features will also be added, such as further collaborative components and translation support. Some findings of the user studies were not conclusive, e.g. preference of result list presentation style and suggestion visualisation. With the evaluation system and scripts in place it will be easy conduct further evaluations on these questions during the following months.

Further evaluations will focus on result surrogate presentations based on the two promising variants, the suggestion dialog as well as the improved and extended collaboration components and will comprise more participants.

5 Conclusion

This deliverable described several user-centred evaluations of interface components conducted after the first phase of development. They showed the general suitability of the developed software. While the evaluations were mostly of a formative nature, the results have been used to improve the usability of the components and to choose between different implementations of interactive features. The user interface is now ready to be evaluated in the context of the complete search system. For these global evaluations, questionnaires and scripts developed during this component evaluation can be re-used or adapted. In addition, the logging framework will allow a complete log of all user activities during the

global evaluations without use of third-party logging software.

6 References

[1] Beckers, Thomas; Dungs, Sebastian; Fuhr, Norbert; Jordan, Matthias; Kriewel, Sascha (2012). ezDL: An Interactive Search and Evaluation System. In: Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval. Department of Computer Science, University of Otago, Dunedin, New Zealand.

[2] Kelly, Diane (2009). Methods for evaluating interactive information retrieval systems with users, Foundations and Trends in Information Retrieval, 3(1-2).

[3] Snyder, Carolyn (2003). Paper Prototyping: The fast and easy way to design and refine user interfaces. In: The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann Publishing, Elsevier, Amsterdam etc.

[4] Raskin, Jef (2000). The Humane Interface: New Directions for Designing Interactive Systems. In: ACM Press Series. Addison-Wesley.

[5] Schaefer, André; Jordan, Matthias; Klas, Claus-Peter & Fuhr, Norbert (2005). Active Support For Query Formulation in Virtual Digital Libraries: A case study with DAFFODIL. In: Rauber, A., Christodoulakis C. & Tjoa, A M. (Hg.), "Research and Advanced Technology for Digital Libraries". Proc. European Conference on Digital Libraries (ECDL 2005), Vol. 3652, Springer-Verlag, pp. 414-425.

[6] Mayring, P. (2003). Qualitative Inhaltsanalyse. Grundlagen und Techniken. 8. Auflage, Weinheim: Beltz, p. 62.

[7] Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In: P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. Usability Evaluation in Industry. London: Taylor and Francis.

[8] Goeuriot, Lorraine; Hanbury, Allan; Jones, Gareth J. F.; Kelly, Liadh; Kriewel, Sascha; Martinez Rodriguez, Ivan; Müller, Henning, Tinte, Miguel A. (2012) - Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System. In Proceedings of Collaborative Resource Development and Delivery.

[9] Gschwandtner, Manfred; Kritz, Marlene; Boyer, Celia (2011), D8.1.2: Requirements of the health professional search. Khresmoi Project public deliverable.

[10] Pletneva, Natalia; Vargas, Alejandro (2011). D8.1.1: Requirements for the general public health search. Khresmoi Project public deliverable.

[11] Chin, J.P., Diehl, V.A., Norman, K.L. (1988) Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. ACM CHI'88 Proceedings, pp. 213-218.

[12] Boyer, Celia; Geschwandtner, Manfred; Hanbury, Allan; Kritz, Marlene; Pletneva, Natalia; Samwald, Matthias; Vargas, Alejandro (2012). D8.2: Use case definition including concrete data requirements. Khresmoi Project public deliverable.

[13] Müller, Henning (2011). D9.1: Report on image use behaviour and requirements. Khresmoi Project public deliverable.

[14] Ignalski, Jessica; Jordan, Matthias; Kriewel, Sascha (2012). Evaluierung von Darstellungs-varianten für Anfragevorschläge bei der Informationssuche. Proceedings of the Workshop "Information Retrieval 2012" (IR-2012), LWA 2012.

[15] Franitza, Markus; Dungs, Sebastian; Kriewel, Sascha (2012). Entwicklung und Evaluierung

einer Rich Internet Application für die Suche nach medizinischer Information. Proceedings of the Workshop “Information Retrieval 2012” (IR-2012), LWA 2012.

[16] Bangor, Aaron; Kortum, Philip; Miller, James (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. In Journal of Usability Studies, Vol. 4, Issue 3, UPA.

[17] Klas, Claus-Peter; Albrechtsen, Hanne; Fuhr, Norbert; Hansen, Preben; Kapidakis, Sarantos; Kovács, Lászlo; Kriewel, Sascha; Micsik, András; Papatheodorou, Christos; Tsakonas, Giannis; Jacob, Elin (2006). A Logging Scheme for Comparative Digital Library Evaluation. In Research and Advanced Technologies for Digital Libraries, Proceedings of the 10th European Conference on Digital Libraries (ECDL). Springer.

[18] Klas, Claus-Peter; Fuhr, Norbert; Kriewel, Sascha; Albrechtsen, Hanne; Tsakonas, Giannis; Kapidakis, Sarantos; Papatheodorou, Christos; Hansen, Preben; Kovács, László; Micsik, András; Jacob, Elin (2006). An experimental framework for comparative digital library evaluation: the logging scheme. In JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries. ACM Press, New York, NY, USA.

[19] Beckers, Thomas; Korbar, Dennis (2011). Using Eye-Tracking for the Evaluation of Interactive Information Retrieval. In: Comparative Evaluation of Focused Retrieval - 9th International Workshop of the Initiative for the Evaluation of XML Retrieval, pp. 236-240, Springer.

7 Appendixes

7.1 Evaluation Script

The script for conducting the evaluation is based on suggestions in [3] and uses parts of the "Sample Informed Consent Form" presented therein. This script or translations of it were used by the evaluation facilitator to conduct experiments B1, B2, B3, and C.

Start session by greeting participant

"Hello, *XXX*. My name is *XXX*. Thank you very much for agreeing to participate in this study. *(This is my colleague, XXX)*"

Possibly try to establish some rapport with participant by doing small-talk and asking about their background.

Brief the participant about the study

"This is a usability study about a new search and information management interface we are developing in an EU funded project named Khresmoi. The interface is intended for people searching for medical and health-related information. Our goal is to make the interaction as appealing, user-friendly and intuitive as possible, and your participation will help us with this.

In this study you will be working with several prototypes of different features for the search interface that we're developing. I will ask you to perform a number of tasks that we believe, people using this interface will typically want to perform, such as running a search or sharing a document with other users. During this, I *(and my colleague)* will be sitting here in the room, observing the session and taking notes. If you are truly stuck or have questions, *(I/my colleague)* will be available. At various points we may ask *you* some questions, e.g. why you used a specific function.

Currently all the documents are in English. Do you feel comfortable searching for English documents?"

"*(During – some of – the tasks, an eye tracking system and/or a web-cam will record the experiment."*

Explain exact nature of recording. Explain eye-tracking if necessary.

"You will also be asked to fill out one or more questionnaires about the software you are testing. All the data collected during this session will be confidential and will only be used in an anonymised fashion for purposes of our research within the project. We may publish results from this and other sessions, but the observations will be anonymised and no publication will include your name.

A very important note before we start: we are testing our software, we are not testing you! This is an early version of the interface and we want to find out which aspects of the software are confusing or do not work. If you have any problems during the tasks, they will very likely be the fault of the interface and discovering those problems will help us make it better.

To the best of our knowledge there are no risks associated with participating in this study. We will *(pay you 10 Euro at the end of this session/certify the participation)* at the end of this session, which will probably last about *(30 minutes/an hour/90 minutes)*. If you need to take a break that's fine, and you may stop your participation in this study at any time without giving a reason. You will be paid regardless."

Obtain the informed consent

"Before we start, I'll need you to sign this document. It summarizes what I just told you. Please read it and if it's okay sign here..."

Explain testing equipment and/or protocol

For "normal" experiment: Start the program for the participant.

"This is the experiment computer. I've already started the program. I will now give you a description of the tasks. Please read them and ask if any part of the tasks is unclear to you. As soon as you are ready and comfortable, use your mouse to click on this button. This will start the task. Do the task at your normal pace, do not try to hurry. If you think you are finished, please tell me."

For an eye-tracking experiment:

"This is our eye-tracking system. Before we start the task we need to calibrate the machine, so that we can track were you are looking at...."

At this point some technical explanations will be necessary. After that continue with normal task protocol.

For multiple experiments during one session it may be necessary to start the program and/or re-calibrate the eye-tracking system before each experiment.

Give task description

Either present a written description of the task or explain the task verbally to the participant.

If user becomes stuck during the session

Snyder [3] recommends to resist the temptation of answering questions that users have: "Once you explain something to a user, you forever lose an opportunity to understand the problem." However, if the user is truly stuck it is necessary to help them over the hump so that the experiment can progress to the next task. A helpful way to do this, without telling the user what to do, is asking questions that lead the user to their goal. From [3]:

1. What are you trying to do right now?
2. What do you think should be the next step?
3. Do you see anything that might help you?
4. *[hint to appropriate interface element]*

End the session and thank the participant

"That's it! Thank you very much, this was a great help."

Give them their money, and have them sign a receipt. Ask them if they have any remaining questions. Escort them out.

7.2 Tasks for experiment B1

In this experiment we are looking at the usability of an interface for searching and for working with search results. Please read the following tasks and ask if any of them are unclear to you.

Please try to complete all the tasks without help. If you are really stuck, you can ask the facilitator.

After the tasks you will be given a questionnaire.

1. Use the program to search for "aspirin".
2. Within the search results find the latest (newest) publications about the topic.
3. Filter the existing results to find publications that are also about "asthma".
4. Remove the filter.
5. Use the program to search for "health diet".
6. Within the search results find the most frequent word (aside from "health" and "diet") using the functionality of the program. Then search for that word.
7. Look at the preview for the three best results of your last search.
8. Go back to the first preview.
9. Open the original page for that result in a browser.
10. Go back to the program. Within the preview, you'll see a few image thumbnails for the current result document. Open a full-size version of the image.
11. Now search for more, similar images based on the image you're currently looking at.
12. The result presentation will have changed to a grid-based look. Switch back to the original list look.
13. Look for a list of your previous searches and go back to the results of your very first search.
14. Change the result display, so that it is grouped by the decade of publication.
15. Find the decade with the most publications.

7.3 Tasks for experiment C

In this experiment we are looking at the usability of an interface for organizing and sharing found documents, and for managing a user profile. Please read the following tasks and ask if any of them are unclear to you.

Please try to complete all the tasks without help. If you are really stuck, you can ask the facilitator.

After the tasks you will be given a questionnaire.

1. Use the program to search for "lung cancer".
2. Save the documents with the titles *"German cancer statistics 2004"* and *"Primary lung tumour visualised by transthoracic echocardiography"*, as well as two documents of your own choice to your personal document collection.
3. Find the documents you just saved in the program and add the tag "lung cancer" (a short descriptive label).
4. Remove the tag from the document "*German cancer statistics 2004"*.
5. Share the document with the title "*Genomic imprinting and assisted reproduction"* with users Max Mustermann and Maria Musterfrau.
6. Create a new, private group with the name "Interesting Documents" to make future sharing of documents easier.
7. Add the two users Max Mustermann and Maria Musterfrau to the group.
8. Use the search to find a user with the name "Hans Maulwurf".
9. Look at the user's profile.
10. Add the user to the newly created group.
11. Go back to your personal document collection and try to find a document that you saved last month. Look at the preview.
12. Now find a document by author "M. Cieslak". Look at the preview and try to print it.
13. Find those documents tagged with "unimportant" and delete them from your collection.
14. Look for a way to edit your own user profile and add a short description.

7.4 Tasks for experiment D

In this experiment we are looking at the usability of a new web interface for searching health information. Please read the following tasks and ask if any of them are unclear to you. You have a maximum of 10 minutes to complete each subtask.

1. Search for documents on “diabetes”
    1. Group the results by type and sort by relevance (best first)
    2. View definitions for two of the result terms
    3. Copy one of the definitions to your tray
    4. Filter the results for clinical studies in New York and print one of them
    5. Ungroup the results and sort by reverse relevance
    6. Remove the filter and find a recent (2011) document about kids with diabetes, copy it to your tray
    7. Search for additional documents on the topic and also copy them to the tray
    8. Save the results from the tray
2. Search for a health-related topic of your choice, using the term suggestions presented at query formulation time. Copy interesting documents to the tray.
3. Search for documents on the human “heart” and use the result images to identify three parts of the heart. Copy the image that was most helpful to answer this question to your tray.
4. Search for “nitrofurantoin” and find the halflife of this drug.
5. Use the query history to repeat your query for “heart”.
6. Search for documents on “smoking”
    1. Filter for documents by the author “Reiner Hanewinkel”
    2. View the details and copy one document to the tray
    3. Group by type of document and sort by year (oldest first)
    4. Remove the filter and search for a document on women smoking after giving birth from 2006, copy that document to tray
    5. Search for additional documents on the topic and also copy them to the tray
    6. Print one of those documents from the tray
    7. Ungroup the results and sort by relevance (best first)
    8. Save the results from the tray

7.5 Questionnaires

The following questionnaires (or German language versions of the same questionnaires) were used during the experiments:

- a (demographic) pre-experiment questionnaire
- the SUS usability questionnaire
- a usability questionnaire based on QUIS
- a short questionnaire on the result item variations
- a short questionnaire on the result list variations
- questionnaire from experiment D

Experiment D Questionnaire

	1	2	3	4	5
1. The software was designed appropriately to allow searching information					
2. The control elements are ordered in an intuitive fashion					
3. The meaning of icons was immediately clear					
4. The usage of filters was obvious					
5. The font size was too small					
6. The layout of the web interface was appropriate					
7. The layout of the search tool was uncluttered					
8. The layout of the tray tool was cluttered					
9. It was possible to smoothly interact with the interface					
10. The layout of the detail tool was uncluttered					
11. The result view was uncluttered					
12. The usage of the grouping function was not obvious					
13. The detail tool allowed for sufficient interactions					
14. The web interface sometimes reacted too slowly					
15. The information presented in the result view was insufficient					
16. The animations did not distract					
17. The usage of the sort function was not obvious					
18. The tools behaved in an expected way					
19. The web interface was tolerant against user errors					
20. Input errors were easily corrected					
21. The font size was to large					

22. The control elements in the search tool were arranged in a confusing way					
23. The visualisation of the process was helpful					
24. The control elements in the tray tool were clearly arranged					
25. The control elements in the detail tool were arranged in a confusing way					

	1	2	3	4	5
1. The web interface quickly allowed a successful search					
2. There were no unnecessary steps in the search process					
3. The software was designed appropriately to allow searching information					
4. The control elements were arranged in an intuitive fashion					
5. The meaning of icons was initially confusing					
6. The usage of the filter function was obvious					
7. The font size was too small					
8. The layout of the web interface was appropriate					
9. The layout of the search tool was cluttered					
10. The web interface allowed only sensible operations					
11. The layout of the tray tool was cluttered					
12. It was possible to smoothly interact with the interface					
13. The layout of the detail tool was cluttered					
14. The layout of the result view was clear and uncluttered					
15. The web interface supports cancelling a search action					
16. The usage of the grouping function was not obvious					
17. The layout of the query history was cluttered					

18. The detail tool allowed for sufficient interactions					
19. The web interface sometimes reacted too slowly					
20. The information presented in the result view was insufficient					
21. The search suggestions of the web interface were helpful					
22. The layout of the search suggestions was appropriate					
23. The animations did not distract					
24. The usage of the sorting function was not obvious					
25. The tools behaved as expected					
26. The web interface did not provide sufficient feedback					
27. The web interface sometimes provided false feedback					
28. The web interface was tolerant against user errors					
29. Input errors were easily corrected					
30. The web interface offered sufficient ways to personalise it					
31. The font size was to large					
32. The search tool was arranged in a confusing way					
33. The query history was helpful					
34. The visualisation of the process was helpful					
35. The control elements in the tray tool were clearly arranged					
36. The control elements in the detail tool were arranged in a confusing way					
37. The control elements in the query history were clearly arranged					

Pre-experiment Questionnaire

Age _____ **Gender** _____

Native language _____

Highest level of education _____

Language skills

	Can easily comprehend simple issues	Can easily comprehend complex issues	Native speaker or comparable to a native speaker
_____ skills			
English skills			

Computer skills

	Never 1	2	3	Daily 4
Do you use a computer for private tasks?				
Do you use a computer for job or education related tasks?				
Do you search the WWW for web pages?				
Do you search the WWW for special types of documents or material (e.g. PDFs, images, definitions)?				
Do you search the WWW for health related information?				

What type of computer do you use, if any? (e.g., Windows-Desktop, Mac, Tablet-PC)

_____

_____

Use of programs or web application

	Have used before	Use regularly
ezDL		
Google search		
Bing search		
Google image search or other image search engine _____		
Other web search engine _____		
Search function of Windows, MacOS, etc.		
Windows Explorer, Apple Finder or other desktop file manager		
Mendeley		
Bibsonomy		
An online library catalog		
Facebook		
Google+		
Other Social Network with personal user profile _____		
Other Social Network with self defined contact groups _____		
Other Social Network that allows *Sharing* of documents _____		

Usability questionnaire

Usability of the Software (SUS)

	strong reject				*strong accept*
	1	2	3	4	5
Ich denke, ich würde dieses System häufig benutzen wollen. *I think that I would like to use this system frequently.*					
Ich fand das System unnötig komplex. *I found the system unnecessarily complex.*					
Ich fand, das System war einfach zu benutzen. *I thought the system was easy to use.*					
Ich denke, dass ich die Hilfe eines Technikers brauchen würde, um dieses System benutzen zu können. *I think that I would need the support of a technical person to be able to use this system.*					
Ich finde die unterschiedlichen Funktionen des Systems sinnvoll integriert. *I found the various functions in this system were well integrated*					
Ich denke, das System enthielt zu viele Inkonsistenzen. *I thought there was too much inconsistency in this system.*					
Ich glaube, dass die meisten Leute die Verwendung des Systems schnell lernen könnten. *I would imagine that most people would learn to use this system very quickly.*					
Ich fand, dass das System sehr umständlich zu benutzen ist. *I found the system very cumbersome to use.*					
Ich fühlte mich sicher im Umgang mit dem System. *I felt very confident using the system.*					
Ich musste einiges lernen, bevor ich das System wirklich nutzen konnte. *I needed to learn a lot of things before I could get going with this system*					

Bildschirmdarstellung (Screen)			
Zeichen am Bildschirm lesbar *Reading characters on the screen*	schwierig *difficult*	○ ○ ○ ○ ○	einfach *easy*
Hervorhebungen vereinfachen Aufgabe *Highlighting simplifies task*	niemals *never*	○ ○ ○ ○ ○	immer *always*
Bezeichnungen und Systeminformation (Terminology and System information)			
Verwendung von Bezeichnungen im System *Use of terms throughout system*	inkonsistent *inconsistent*	○ ○ ○ ○ ○	konsistent *consistent*
Begriffe sind passend zu Aufgaben *Terminology related to task*	niemals *never*	○ ○ ○ ○ ○	immer *always*
Position von Benachrichtigungen auf Bildschirm *Position of messages on screen*	inkonsistent *inconsistent*	○ ○ ○ ○ ○	konsistent *consistent*
Eingabeaufforderungen *Prompts for input*	verwirrend *confusing*	○ ○ ○ ○ ○	klar *clear*
Programm informiert über Fortschritt *Program informs about its progress*	niemals *never*	○ ○ ○ ○ ○	immer *always*
Fehlermeldungen *Error messages*	nicht hilfreich *unhelpful*	○ ○ ○ ○ ○	hilfreich *helpful*
Erlernbarkeit (Learning)			
Neue Funktionen praktisch auszuprobieren *Exploring new features by trial and error*	schwierig *difficult*	○ ○ ○ ○ ○	einfach *easy*
Es ist klar, wie Aufgaben zu erledigen sind *Performing tasks is straightfoward*	niemals *never*	○ ○ ○ ○ ○	immer *always*
Systemfähigkeiten (System Capabilities)			
Systemgeschwindigkeit *System speed*	zu langsam too slow	○ ○ ○ ○ ○	schnell genug fast enough
Zuverlässigkeit *System reliability*	unzuverlässig unreliable	○ ○ ○ ○ ○	zuverlässig reliable
Fehler korrigieren *Correcting your mistakes*	schwierig *difficult*	○ ○ ○ ○ ○	einfach *easy*
Für jedes Erfahrungsniveau geeignet *Designed for all levels of user*	niemals *never*	○ ○ ○ ○ ○	immer *always*

Questionnaire on result item variations

	Strong Reject				*Strong Accept*
	1	2	3	4	5
To judge the relevance of result items one of the representations was clearly better than the rest					
I thought that it depended on the search task, which representation was best.					

To decide which of the results was relevant for the search task,	*Strong Reject*				*Strong Accept*
	1	2	3	4	5
Title of document (1)					
Document rank (2)					
Relevance bar (3)					
Publication date (4)					
Publisher / Organisation (5)					
Authors (6)					
Text snippet (7)					
Thumbnails (8)					
Target audience (9)					
Other classification labels (10)					
URL/Link (11)					
Ratings of other users (12)					
Highlighting of query terms (13)					
Color highlighting of different parts (14)					

Additional questions:

- Which parts of the result representation were unclear (you did not know what information was shown)?
- Overall, which of the four variants did you prefer?

Questionnaire on result list variations

	Strong Reject				*Strong Accept*
	1	2	3	4	5
To find exactly the result I was looking for, the grouped presentation of the result list was helpful.					
To find exactly the result I was looking for, the result presentation in separate tabs was helpful.					

Additional question:

- Overall, which of the two variants did you prefer?

7.6 Logging Events

In the following section all log events that were captured for the usability and component evaluations are listed. The log events cover all possible user interactions with the search interface that actuate changes in the system. Interactions that do not change the state of the system, such as mere mouse movements, or opening/closing of menus without selection were not logged – but are visible on the screen recordings.

The log events are grouped by interface part and list the user action and the logged parameters. For example, if the user adds a new search term to the query, the system will log the query field that was changed (e.g. “author”, “title” or “all”), the previous query and the query after the addition. Not listed but logged for all events are a user id, a unique session id, a timestamp, and a sequence number within the user's session.

Search form

- User enters a new search term: field, previous query, new query
- User adds a new positive/negative example image to query: image url, pos/neg
- User changes example image to positive/negative: image url, pos/neg
- User changes an existing search term: field, previous query, new query
- User deletes a search term: field, previous query, new query
- User deletes an example image from query: image url
- User clears the query form
- User sends the query: query, constraints
- User drops and combines a query from query history: previous query, dropped query, chosen combination option
- User triggers the suggestion dialog: token for which suggestion dialog was triggered, content of dialog
- User selects a suggestion without changing query: position of suggestion in list, type of suggestion (spelling, translation, etc.), content of suggestion?
- User chooses a suggestion to modify query: position of suggestion in list, replaced token, suggested token
- User dismisses suggestion dialog

Result view

- User cancels query before it finishes
- User chooses to see results now before query finishes: list of digital libraries with answers, number of documents already found
- User filters results: filter terms, number documents before and after filtering
- User removes result filter
- User switches the result presentation to grid/list: new presentation
- User changes the sorting criterion/direction: new criterion, new direction
- User extracts terms/years/authors/libraries from results: attribute chosen
- User exports result item(s): result item(s) exported, position in result list, format (BibTeX, Text, HTML, ...)
- User activates/deactivates group explorer window: activation or deactivation
- User chooses to hide all/show all of a result groups' contents: hiding or showing
- User hides/shows a single result group's contents: hiding or showing, group name/identifier
- User changes the option to group by: new grouping option
- User selects or de-selects result item(s) with mouse or keyboard: result item(s) selected, position(s) in result list, selection or de-selection

- User selects or de-selects all items in group: group, selection or de-selection
- User scrolls the result list: direction
- User opens a result item in new tab: result item, position in result list
- User opens a result item in new window: result item, position in result list
- User copies current document to tray: result item, position in result list
- User copies current document to personal library: result item, position in result list

Detail view

- User looks at details for a result item: item
- User looks at details for a tray item: item
- User looks at details for an item from personal library: item
- User copies current document to tray: item
- User copies current document to system clipboard: item
- User copies current document to personal library: item
- User exports current document: item, format
- User prints current document: item
- User adds user of currently view profile to existing groups: list group ids, user id
- User goes to previous detail view: item
- User goes to next detail view: item
- User clicks on a search link from details: query field, search terms
- User clicks on similar images link from details: image url
- User views full size version of image: image url
- User clicks on an external or full-text link: link url

Query history

- User filters queries: filter terms, number queries before and after filtering
- User removes query filter
- User un-/groups by date: grouped or ungrouped
- User activates/deactivates group explorer window: activation or deactivation
- User chooses to hide all/show all of a result groups' contents: hiding or showing
- User hides/shows a single result group's contents: hiding or showing, group name/identifier
- User selects or de-selects query/queries with mouse or keyboard: result item(s) selected, position(s) in result list, selection or de-selection
- User selects or de-selects all items in group: group, selection or de-selection
- User sets a historic query: query, position in list
- User sets and executes query: query, position in list
- User deletes query from history: query, position in list
- User scrolls query list: direction

Personal library

- User drops and adds item(s) to personal library: item dropped
- User filters library: filter terms, number items before and after filtering
- User removes filter
- User un-/groups by date: grouped or ungrouped
- User activates/deactivates group explorer window: activation or deactivation
- User chooses to hide all/show all of a result groups' contents: hiding or showing
- User hides/shows a single result group's contents: hiding or showing, group name/identifier
- User selects or de-selects item(s) with mouse or keyboard: result item(s) selected, position(s) in result list, selection or de-selection
- User selects or de-selects all items in group: group, selection or de-selection

- User exports item(s): result item(s) exported, position in result list, format
- User scrolls item list: direction
- User deletes item from personal library: item, position in list
- User adds a tag to item: item, tag
- User removes a tag from item: item, tag
- User removes all tags from item: item
- User shares a document with other users: item, list of user ids
- User cancels sharing of a shared document: item, list of user ids

Extraction view

- User switches display of extracted terms to cloud/list: extracted attribute displayed, new presentation
- User selects a term from extraction view: term, count, presentation used
- User sets and executes a query with a term: term, count, presentation used

Group creation and management

- User creates a new group: group id, name, privacy status
- User renames group: group id, old name, new name
- User changes privacy status: group id, new status
- User switches display between all visible and owned groups: new display
- User opens details for a group: group id
- User deletes a group: group id, number members
- User joins a (public) group: group id, number members
- User leaves a (public) group: group id, number members
- User adds a new member to a group she owns: group id, user id
- User removes a member from a group she owns: group id, user id

Library Choice

- will not be logged for WP3 evaluation since libraries used will be fixed

Tray

- User drops and adds item(s) to personal library: item dropped
- User filters tray: filter terms, number items before and after filtering
- User removes filter
- User un-/groups by date: grouped or ungrouped
- User activates/deactivates group explorer window: activation or deactivation
- User chooses to hide all/show all of a result groups' contents: hiding or showing
- User hides/shows a single result group's contents: hiding or showing, group name/identifier
- User selects or de-selects item(s) with mouse or keyboard: result item(s) selected, position(s) in result list, selection or de-selection
- User selects or de-selects all items in group: group, selection or de-selection
- User exports item(s): result item(s) exported, position in result list, format
- User scrolls item list: direction
- User deletes item from personal library: item, position in list
- User shows or hides creation panel: show or hide

- User creates a new term: term created
- User creates a new author: author created

Perspectives and Tools

- User moves a view to new position: view, old and new position
- User closes a view: view, old position
- User opens a new view: view, new position
- User maximizes view: view
- User undocks/docks view: view, docking or undocking
- User brings open view to front: view
- User switches to new perspective: old and new perspective
- User saves perspective: file name
- User resets current perspective: perspective
- User removes custom perspective: file name

Desktop

- User starts desktop/application: login name
- User exits desktop
- User resizes desktop: old dimensions, new dimensions
- User selects "about dialog"
- User opens help
- User imports document from bibtex file: file name or URL, tool to import to
- User opens the "option dialog"

7.7 Gaze Plots and Heat Maps

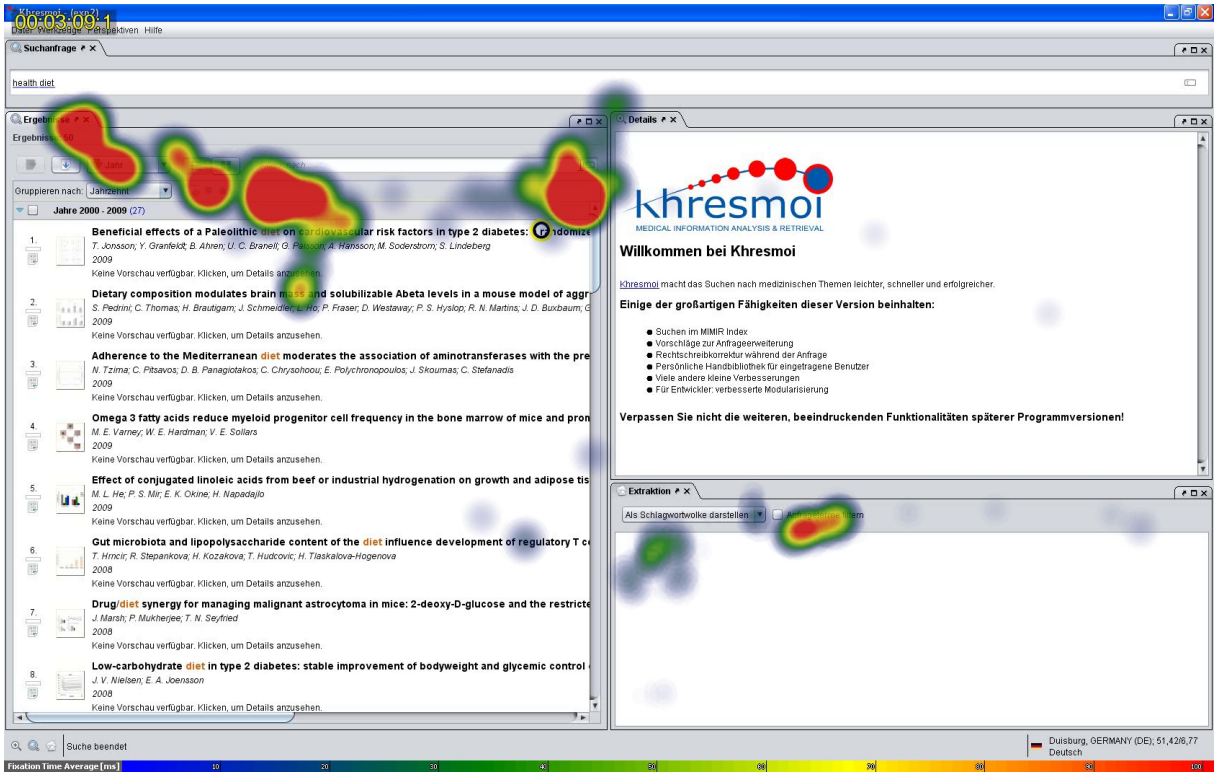


Figure 20: Heat Map of a user searching for the extraction tool

Figure 21: Heat Map of a user searching for the switch between grid and list view

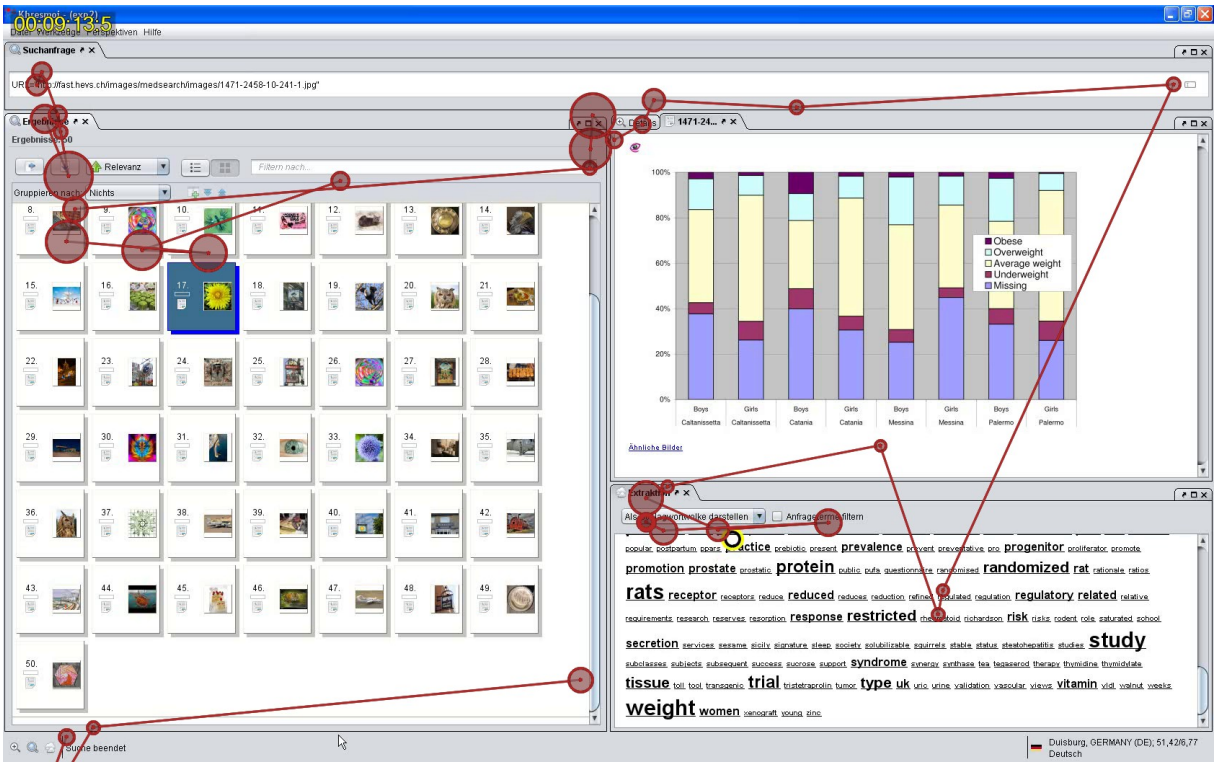


Figure 22: Scan Path of a user searching for the switch between grid and list view

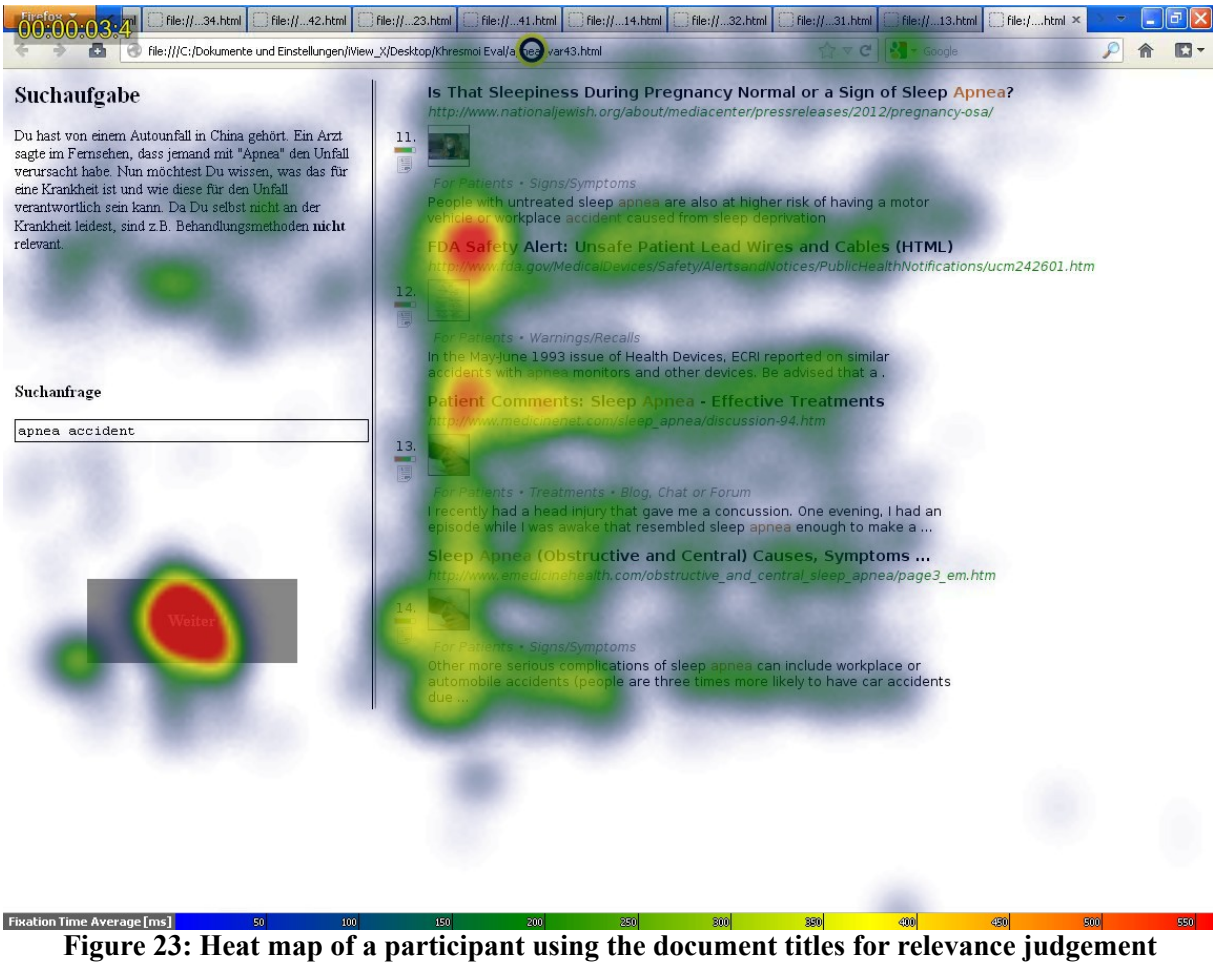


Figure 23: Heat map of a participant using the document titles for relevance judgement

Figure 24: Heat map of a participant neglecting the thumbnails

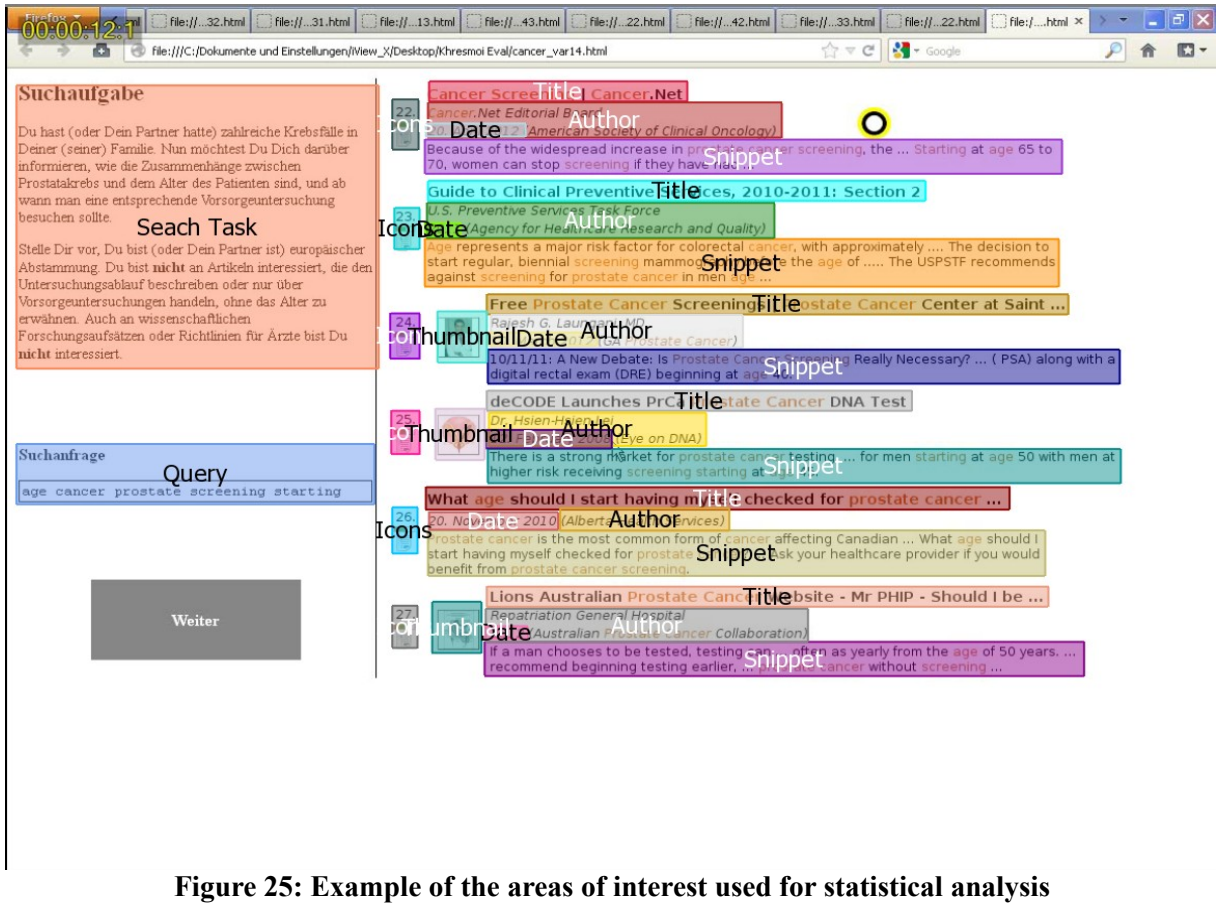


Figure 25: Example of the areas of interest used for statistical analysis