Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

D3.7: Report on results of the WP3 second evaluation phase

Deliverable number	*D3.7*
Dissemination level	*Public*
Delivery data	*08.07.2014*
Status	*Final*
Authors	*Thomas Beckers, Tina Bannert, Sebastian Dungs, Matthias Jordan, Noel Kamda, Sascha Kriewel, Andreas Tacke*

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.

Executive Summary

In this deliverable we present the results of the second user interface component evaluation phase. Several user–centered experiments were conducted, some of them using eye–tracking equipment. In addition, a number of bachelor, master and diploma theses completed at the University of Duisburg–Essen examined the effect of various user interface and user support features.

The usability experiments for the search and collaboration reported in D3.2 (Report on results of the WP3 first evaluation phase) were repeated with updated interface component versions and a larger sample size (20 participants instead of 8). They show that the changes introduced since the first evaluation resulted in an improved user experience. The usability evaluation showed improvements for the subjective System Usability Scale (SUS) score as well as for error rates for most tasks.

Table of Contents

Abbreviations

ezDL	Easy Access to Digital Libraries
GAW	Society of Physicians in Vienna
GUI	Graphical User Interface
KP	Khresmoi professional
QUIS	Questionnaire for User Interface Satisfaction
SUS	System Usability Scale
UDE	University of Duisburg-Essen
TCR	Task Completion Rate
TUW	Technical University Vienna
UI	User Interface
WP	Workpackage
WP3	Workpackage 3 of Khresmoi (User Interface components)

1 Introduction

This document describes the outcomes of the second and final user–centred evaluation of the user interface components carried out in task 3.6. Several experiments have been conducted during the second half of the project, each focussing on different aspects of the user interface. All experiments took place in a lab setting while different means of data acquisition were used, such as eye–tracking, screen capturing, logging of user actions and events as well as questionnaires. Experiments were designed to be comparable to the results given in D3.2 (Report on results of the WP3 first evaluation phase) while addressing recommendations given in D7.2 (Meta-analysis of the first phase of empirical and user-centred evaluations). Therefore, a larger sample size (20) was used for the main parts of the second evaluation, as was suggested in D7.2. Additional evaluation metrics such as task–specific error rates were also included in the evaluation and compared to figures compiled from the analysed data of the first evaluation phase.

One experiment focused on the collaborative components. Following the recommendation given in D7.2, it was designed to verify and expand the findings of D3.2, which were based on a very early version of the collaborative interface components. In addition the SUS score, average task completion rate and task–specific error rates are reported. Due to the lack of a gold standard or other comparable systems, usability scores and error rates were compared to the results from the first evaluation phase. Experiments evaluating the effect of user–specific information on ranking are beyond the scope of this deliverable which is to describe the evaluations of user interface components developed in WP3. Adhoc tests of different ranking settings using the users' language, country, and medical knowledge were carried out at TUW and GAW. They indicate that using these aspects profiling information can have a positive effect on the ranking quality. However, for larger document collections the negative effect on retrieval speed made their use impractical for the current prototype.

The remainder of this document is structured as follows: Section 2 gives an update on the usability evaluations from D3.2 which were repeated with a larger sample size. Next, the results of the evaluation of suggestions and search refinement components are specified in Section 3. Section 4 covers user input for updating resources (specifically translations) while Section 5 focuses on the result presentation. Findings concerning the collaborative components and the personal library are detailed in Section 6. Section 7 concludes the document.

2 Update on usability evaluation from D3.2

As part of deliverable D3.2 a small scale, formative usability study on the basic functionalities of the Khresmoi user interface components was conducted (cf. Section 3.2.1 of [2]). It was a task–driven study with eight participants aimed at discovering usability problems. The tasks covered searching, filtering, grouping and sorting results, extraction of common keywords, working with the preview and query history.

The study in 2012 resulted in a System Usability Scale (SUS) score of 66.9, slightly below the mean score of 68 [12] or 69.5 [1] for SUS studies. This was intended to be used as a baseline for another usability study at the end of component development. The answers to the SUS questionnaire were used to drive development of an improved user interface.

During the month of June 2014 a second usability study was conducted at the University of Duisburg–Essen. A total of 20 participants took part in the experiment, each evaluation session taking roughly an hour (the experiment was combined with the usability study described in Section 6). Where possible, equivalent tasks to those given during the previous study were used, with the aim of having the participants discover and use all parts of the user interface. The tasks were provided in German. An English translation of the tasks is provided in Appendix A.

Eleven of the 20 participants were female, most were students of various subjects including teaching, social science, communication science, psychology, environmental engineering and computer science (ages ranged from 21 to 34, with an average age of 25). Five had already obtained a bachelor's degree. All but three participants used the German version of the interface. As during the previous study, the participants were asked about their previous experience with computers and web search systems. All participants used a computer and web search engines often or daily. None had any experience with Khresmoi or the ezDL software.

The sessions were captured with screen recording software and an eye–tracker was used to record gaze positions. Not all eye-tracking sessions were usable for analysis, as sometimes e.g. glasses or eye make–up can prevent accurate capturing of eye movements. A total of 12 hours and 12 minutes of video was recorded (on average 38.5 minutes of video per participant, not counting time in between experiments) and analysed. As before, an SUS questionnaire was given to the participants after all tasks were completed. By adding up the answers (recoded as 0-4) and multiplying by 2.5, a score between 0 and 100 can be computed. This resulted in an average score of 72.1 for the complete questionnaire (with a standard deviation of 8.6), a clear improvement compared to the formative evaluation results and a "good" score according to [1].

As shown by Lewis and Sauro using factor analysis, the SUS score can be decomposed into two separate scores for learnability (2 items) and usability (8 items) [10]. This shows an average value of 81.7 for learnability and a value of 69.7 for usability, suggesting that the system can be quickly learned despite its comprehensive functionality, but that it can still be improved regarding its usability.

2.1 Individual Tasks

The screen recordings were analysed and combined with logs and notes from the experiment facilitators to gain more detailed insights into remaining usability issues and areas that could benefit from additional affordances. In the following we look at individual tasks that provided interesting results. The average combined task completion rate (TCR) for all users and tasks

was .752 (or 75.2%) with a .107 standard deviation. Several complex or difficult tasks (see below) resulted in a combined TCR that was slightly below average (according to Sauro [13]). For standard tasks such as searching and filtering, calculated with the Adjusted Wald method [11], the confidence interval for expected TCR is between .858 and 1 (with alpha =.05) and the best estimate is 95.4%. For the most difficult tasks, such as determining the level of readability, the confidence interval for TCR is between .342 and .742 with the best estimate at 54.5%.

Task specific error rates have been compiled in Table 1 and compared with the error rates from the first evaluation phase where applicable. Tasks using functionality that was only added after the first evaluations have no corresponding task in D3.2 (experiment B1). Most error rates improved (i.e. decreased). The image search task was more difficult during this evaluation than during the previous one, which may explain the increase in problems observed.

Sort or group by date

When asked to find the newest results for the given query, successful users showed two basic strategies. The first group immediately looked (and found) the extended search options and switched from sorting by relevance to sorting by recency. The second group used the facet browser to restrict the date facet to the last three month. Then they either scanned the remaining documents for the most recent one, or started searching for a way to switch sorting. Unsuccessful users either tried to scan the entire result list for recent documents, restricted to the last three months but stopped there, or (falsely) identified the documents listed in the "news" category as the most recent ones.

Eleven users were completely successful, seven were partially successful (restricting the date facet, but not realizing that the results are still sorted by relevance) and two users were unsuccessful in solving the task.

Task	D3.2	Error rate	D3.7	Error rate
Search	1+5	.062	1+6	.000
Sorting	2	.375	1.a	.275
Readability	—	—	1.b	.450
Details	7	.125	2	.125
Go Back	8	1	2.a	.425
Browser	9	.125	2.b	.175
Filter	3	.000	3+4	.000
Adv. Search	—	—	5+8	.277
Switching	12	.250	6.a+c	.037
Image Search	11	.000	6.b	.250
Search history	13	.250	7	.150

Table 1: Error rates for usability experiment of search tools. For D3.2 and D3.7 the corresponding task numbers are given (see Appendix A).

Document surrogate

Many participants had problems identifying documents (classified as being) written in easy English. Most were able to restrict the results to just English documents, but some then started to look for a filter setting or an additional subcategory among the language facets. Eye-tracking showed that while interacting with the document surrogates some users simply did not see the icons which indicate readability (and two stated so in the after–session questionnaire). Others saw the icons but not the tooltip explaining it (see Figure 1 for the result icons).

Nine users were completely successful, four were partially successful and seven users were unable to complete the task. The suggestion of having the readability classification as a subcategory of the language facet is an option to be explored for future improvements.

Nearly all users were unable to identify particularly trustworthy results using the document surrogates. Only one user was able to solve the task without help. Many did not see the green bar indicating the level of trust the system has in the source as revealed by the eye-tracking (nine users stated so in the after–session questionnaire). Even those who saw the bar usually could not relate it to trustworthiness. Some users suggested different visualization: a thumbs-up icon or a green tick mark. Other users were looking for something similar to the green bar used, but misidentified the relevance bar as a trustworthiness indicator.

Show and interact with details

The interface offers buttons to navigate "back" and "forward" through already visited documents (see Figure 4). However, during the experiment only slightly more than half of the partipants used them while interacting with document details. This is still an improvement over the last evaluation, where none of the participants discovered the navigational buttons. Opening the details themselves or navigating to the original webpage was not a problem for most users. Only two users hesitated to click on an item in the result to show details and no one failed to open the original webpage. Most users chose the link that is provided in the details, while four users tried to double click on an entry in the result list (and opened the webpage in this way).

During the task that asked users to give a definition of a term based on the answers of the system, 16 participants used the automatically generated summary included in the detail view (and were counted as partially successful), while only two used the definition supplied by the system. Two users left the system to look at the original web page for a result (one of them explained that they prefer to look at the formatted full page instead of the summary without formatting).

Use filter

Participants were asked to filter the results so that only documents with the word `Remikiren` are shown, to then remove the restriction, and filter for documents that have both the word `Aliskiren` as well as the word `implication`. Although some users initially misunderstood the wording of the task description, all were able to find the advanced result options and successfully used the filter.

Many users also employed the textual filter field, the multi filter dropdown boxes and the facet filters successfully during other tasks that did not directly call for their use.

A task aimed at participants combining various filter options had to be removed from the experiment since a bug prevented completion. The bug has since been fixed.

Use advanced search

Two tasks prompted participants to use the advanced search features. One task asked to find a page on Wikipedia, while the other specifically asked for definition pages. Both tasks can be easily solved by using restrictions available in the advanced search: results can be restricted by domain or publishing organisation, and a category can be chosen for results. While all

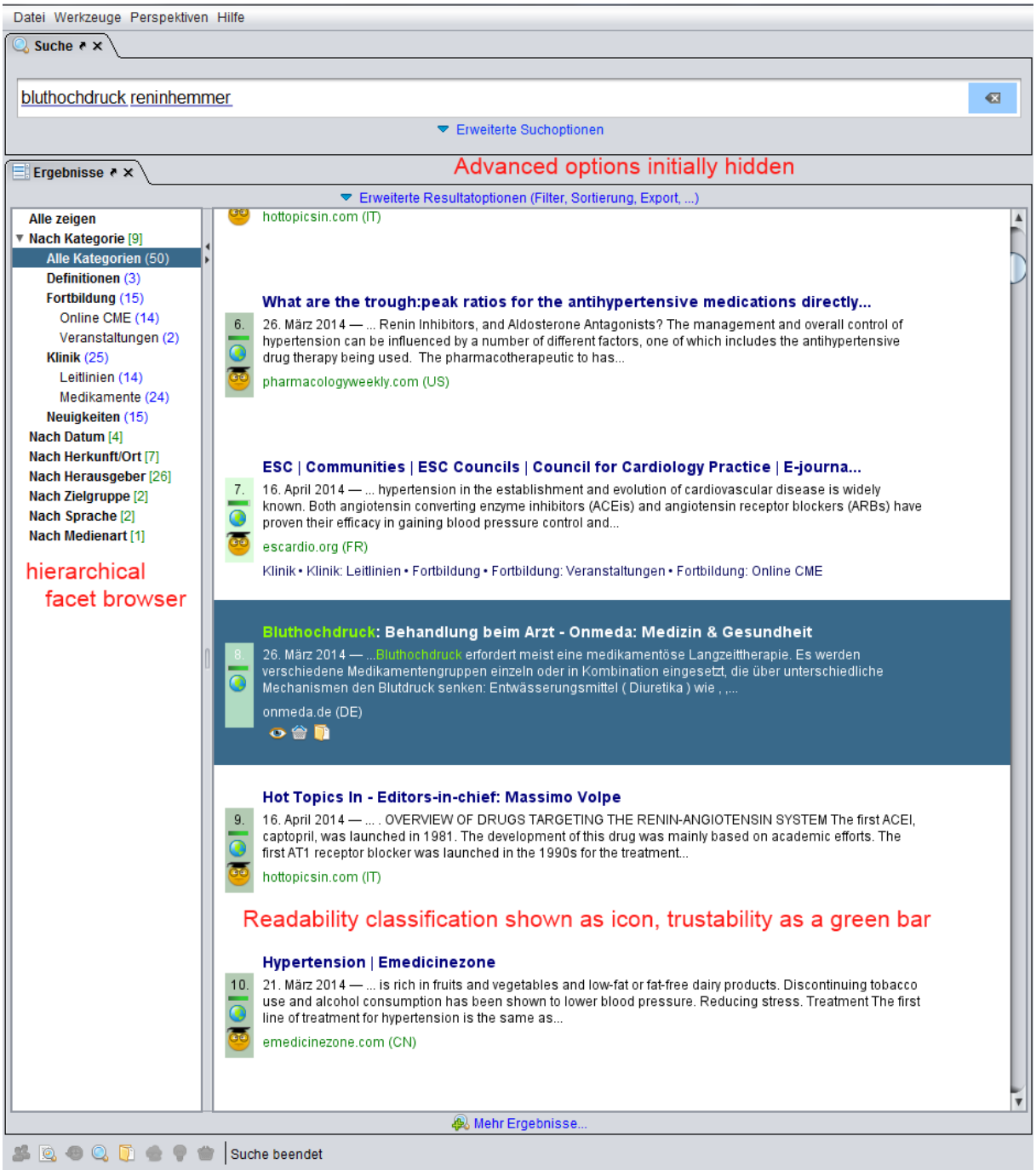

Figure 1: The left side of the Khresmoi Professional desktop client. The interface shows the advanced search and result options initially hidden. Since the last evaluation the facet browser has been changed to a hierarchical presentation. The result items show readability and trustability visualisations.

participants found and opened the advanced search at one point of the experiment, only twelve of 20 used the advanced search for searching only in Wikipedia and only eight of 20 used it to search for definitions only.

Seven more users were successful in searching for Wikipedia pages by specifying the search term `Wikipedia` or `wiki` in the main query field, or by using the facet browser to restrict a more general initial search to just results from Wikipedia. Only one user did not manage to solve the task.

For the second advanced search task, seven participants used the facet browser to restrict a general search result to just definition, one accidentally skipped the task, while four users did not manage to find definitions.

Image search

Switching from web page results to image results was not a problem for nearly all users, with just one user being only partially successful. Most users switched to the image result tab, while five users selected the image media type from the combined result list.

Participants had slightly more problems searching for similar images related to the previous query. Twelve were fully successful, while six users only searched for similar images (without keeping the previous query terms) and then used filters to restrict the results post search. Two users failed the task and were unable to search for similar images.

Search history

Unlike during the previous evaluation, all but three users were able to find the search history without help, although some users spent some time before discovering the menu. No one had problems going back to the results of the first search of the session by either double clicking on the correct history entry, or by selecting it and pressing the "re-run" button.

One of the biggest problems and criticisms during the first evaluation was that newly opened tools are sometimes hidden or open in a region without enough space to fully show them. This was addressed, so that new tools always open in the front and automatically resize the view to be fully visible. If a newly opened view is triggered by a click from another view, the new view will never hide the old view but creates a split window instead. Many users also rearranged tool views to their personal preferences without prompting and were successful in re-docking undocked views (see [3] for a description of the docking framework).

2.2 Summary

In this chapter an update on the usability evaluations from D3.2 was described, using 20 participants and addressing the main criticism from D7.2 – small sample size. The experiment was designed to be comparable with the previous one. Measuring usability with SUS scores, the new version of the interface scored 72.1 points out 100, a small improvement over the first evaluation (66.9). Several tasks were analysed in detail using eye–tracking data and system logs. While some interface features proved to be easily and intuitively understandable (e.g. grouping and sorting of results) others need further improvements. For example, the visualization of reading difficulty and trustworthiness of documents in the result list was not clearly understood by participants. Here additional tooltips or other visualisation options need to be explored. Nevertheless, interaction with the document details as well as the "forward" and "back" mechanism in the detail tool was received positively. Furthermore, filtering and use of the advanced and image search forms was tested successfully. Unlike in the first round of user evaluations, most

of the participants managed to use the program's menu to open additional tools not present in the default perspective.

In conclusion, it became obvious that major parts of the interface were decisively improved over the first interface version. However, there is still room for improvement regarding the visualization of document attributes in the result list.

3 Suggestions of search refinement and continuation options and support

In his diploma thesis Jaroslaw Gustak evaluated the effectiveness of different types of proactive search suggestions [4]. The proactive system was described in deliverable [3]. All modules show their suggestions when the user pauses for a (configurable) amount of time during typing a query. The suggestions are displayed in a popup menu below the current query term. On selecting a suggestion or on a mouse over, an explanation may be shown. For the evaluation three proactive modules were tested:

1. A module which suggests related queries based on the collected query history of the system's user base.

2. A module showing translations for query terms.

3. A module providing spelling suggestions for query terms.

A total of 18 participants took part in the evaluation. Each performed a total of six tasks with predefined queries. For half of the tasks users were supported by the proactive modules, with the combinations of module support and search task being rotated (using a Latin square design).

For measuring the effectiveness of the proactive query specification support, interactive session recall and precision based on the documents saved by the user as relevant were computed. Session recall describes the ratio of relevant documents correctly saved by the user compared to the total number of relevant documents (including the results from query reformulation) in the collection. Session precision describes the ratio of relevant documents correctly identified as relevant by the user compared to the total number of documents saved.

The translation module showed significant improvement versus the baseline without proactive suggestions. Interactive session recall improved by .053 (p=*.018*), while the session precision improved by .084 (p=*.011*). The spelling module also improved session recall and precision, but for both measures there was no significant difference (p=*.695* and p=*.737*). For the suggestion of related queries the module could not provide an improvement versus the baseline performance, instead recall and precision descreased. However, these measures too were not statistically significant.

Looking at the user experience with the proactive interface components, the participants rated the modules between 86.6 (for the query suggestion module) and 91 (for the translation module), which translates to an A- to A grade.

In his diploma thesis Andreas Tacke [14, 15] evaluated the effect of strategic support and tactical suggestions on search success and user experience. Users of search systems often lack the procedural expertise to solve more complex information tasks, even when they have the necessary domain knowledge. The user experiment tested a comprehensive support concept combining a scaffolding approach with a system of situational search suggestions first described by Kriewel [7, 8, 9] and adapted for the project. This concept has been detailed in the previous deliverable [3].

The user experiment is described in detail in [16] and was conducted with 22 participants. The subjects worked on two tasks during which they received tactical search suggestions, the isolated effectiveness of which had been previously shown [8]. Half of the participants were

additionally supported through a system of scaffolds which guided them through the search by dividing a complex task into sub tasks, structuring the process and providing instructions regarding useful tools for each step.

Regarding the efficiency of search, no significant differences were found in the time needed to solve both tasks between the two groups (the group receiving the combined support was slightly faster with 36.3 vs. 35.7 minutes). However, the participants in the scaffolding group used advanced search features significantly more often (p=.*01*) and more frequently requested tactical suggestions. This resulted in a marked improvement of the task completion rate for the group receiving the combined support (with 95% vs. 54%) and a high acceptance rate among the partipants. So, while combining scaffolding with search suggestions does not necessarily help in making a search faster, it can allow searchers to create more effective search strategies and provide a more pleasurable and less frustrating search experience.

4 User input for updating resources

To evaluate the interface components for updating resources the test case of updating automatically translated content was selected. The experiment focused on the user experience during text entry and was conducted as part of the larger usability test reported in Section 2.

Two variant interface components for updating content were created. Even though they were designed to interface with the automatic translation features of the Khresmoi system, they are suitable for use with other textual content.

- *Variant 1* presents a parallel view of the original and the translated text, using the translation judged best by the translation service (see Figure 2). The translated text is editable and the part currently being edited is highlighted in the original text (using the alignment information provided by the service). Suggestions of alternative translations are shown for the current sentence on clicking the right mouse button, but users are free to provide their own translations.

- *Variant 2* presents each sentence of the original text together with the (up to) three best translation alternatives (see Figure 3). For the additional alternatives, green and red annotations are used to show where they deviate from the translation judged best by the service. The entire translated text is shown at the bottom and is updated when users select an alternative. The final translation can be fine tuned by editing the combined text.

During the usability experiment a total of 18 users were asked to use the two translation variants to make updates to the automatic translation for different documents. In a post–experiment questionnaire they were asked to rate the different features on a 5 point Likert scale (encoded from 0 for not at all helpful to 4 for very helpful). Highlighting of the current part of the text was generally recognized as a positive feature with a median of 3 and a mode of 4, while opinions on the offering of translation alternatives was divided with only about half of the users finding them helpful (median 2, mode 3). If translation alternatives are shown, annotations of the differences were appreciated (median and mode 3).

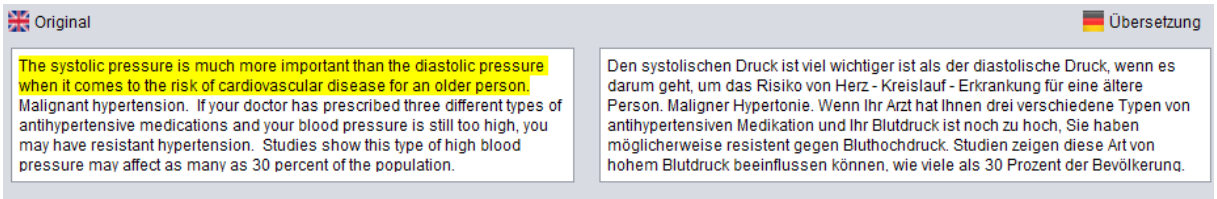

Figure 2: *Variant 1* of the interface component for updating resources. The original text is presented on the left, the editable translation is shown on the right. The corresponding part of the original text is highlighted for the part of the translated text currently being edited.

Analysis of the eye-tracking videos captured during the experiments showed no significant differences in the numbers of errors made during use of the components, nor in the time needed for the update. Most of the time was spend on thinking about the translation improvement, so that any possible speed gain from being able to select an initial translation alternative was lost. Therefore the main difference lies in the user experience during the interaction with the

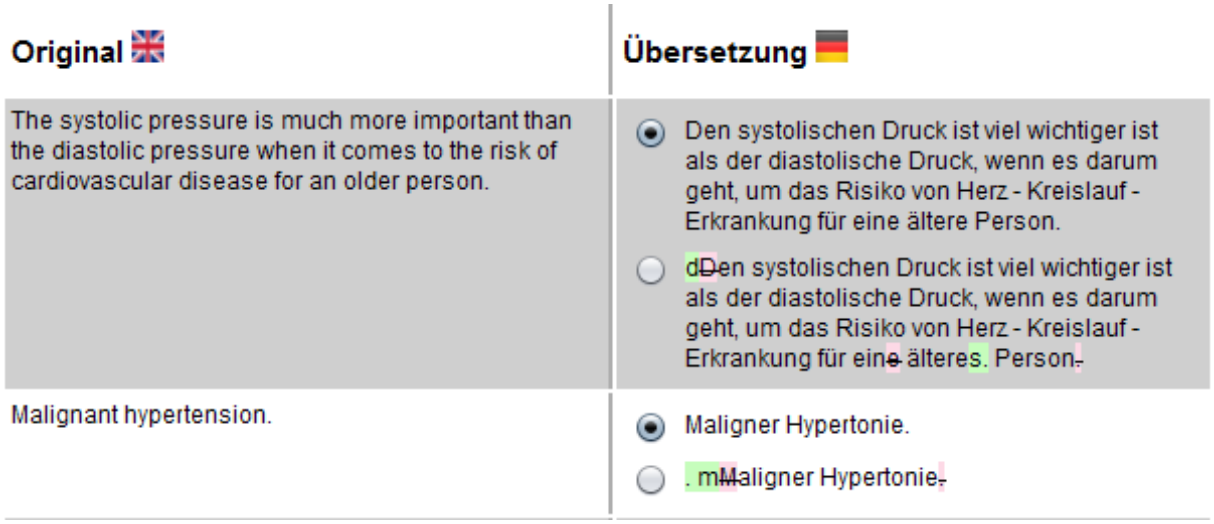

Figure 3: *Variant 2* of the interface component for updating resources. The original text is presented on the left broken into sentences, the translation alternatives are shown on the right. Differences are annotated in green and red.

component. Here the opinions were clearly in favour of variant 1 with the parallel display of the complete text and highlighting to show the text alignment: 10 out of 18 favoured this variant, while only 2 liked variant 2 better, with the remaining participants having no favourite.

In addition, user comments were collected which offer some insights into why they preferred this version.

- Offering multiple options might lead to users "lazily" selecting one without fine-tuning instead of thinking about a better translation (according to two participants). On the other hand, some users had more confidence in the system translations and thought that offering the translation alternatives would lead to less errors. This however depends on the quality of the system translation which was beyond the scope of this deliverable.
- Several users explicitly did not find the sentence level alternatives helpful, and would rather be able read the text in its entirety. They pointed out that the context of the sentence is important for them when updating a translation (e.g. when deciding on demonstratives or reflexive pronouns). The amount of information also required some scrolling by the users which was pointed out as a negative. Still, a few found the second variant to have a clearer presentation.
- The *variant 2* was criticized for offering too much information at the same time. Just showing alternative translation on demand for the current sentence is less overwhelming.
- Most users found the highlighting (based on alignment information) helpful, but one criticized it as distracting and wished for another colour instead of signal yellow.

5 Result presentation

As suggested in the meta–analysis deliverable, additional analysis of the eyetracking data on result presentation from the first evaluation phase was done. In particular the time needed by the participants to make relevance judgements using the different evaluations was extracted from log files and videos and compared. However, no significant results were found.

5.1 Update on result item and result list visualisation

When using the grouped result list variant participants needed on average 221 seconds per task, while using the tabbed result list they needed 203 seconds. Tasks and variants had been rotated during the experiment and both list variants were used to present an equal number of relevant documents. A Welch two sample t-test showed no significance for the difference in time (p=.*42*). While it might be plausible that the separate tabs could make relevance judgements easier and faster, this is not supported by the data collected. Neither were users more successful in separating relevant from non-relevant documents using one or the other of the variants, as reported in the first evaluation deliverable. Eyetracking data from the experiments proved to be too unreliable to reveal any additional insights about skipping behaviour beyond those reported in deliverable D3.2 [2]. Therefore the recommendation given in that deliverable based on user preferences remains.

During the usability experiments described in Sections 2 and 6, the 20 participants were also asked to complete tasks where they needed to interact with the result list. The tasks were specifically designed to confront them with the newly introduced visualisations of trustworthiness and readability. After the experiment the participants were asked to fill out a questionnaire similar to that described in [2] (Section 3.2.3), so that results could be compared.

Of highest importance to users were document title (85% found it important, compared to 81% previously), snippet (85% vs. 87.5%) and publisher (75% vs. 50%). The relevance bar remains relatively unimportant for most users, with nearly half (8) having missed it entirely. However, an equal number found it useful, especially when filtering and re-ordering the result list, as it gave them a good indicator of the result quality independent of position.

Visualizations of trustworthiness of sources and readability were not accepted by the users, even those that saw and understood them during their tasks. Participants stated a reluctance to entrust decisions about these variables to the system, particularly because they found the data to be poor. The system's categorization of readability did not match up with their own perception which led to a rejection of this result item aspect (only 38% positively responded when asked if they found it useful in deciding about selecting a specific result item, even though one of the task in the experiment specifically asked them to select easily readable documents). As discussed above (in Section 2) the trustworthiness visualization was rarely seen or understood, and even after explaining it to participants was rejected by them (only 4 of 20 found it useful).

5.2 Word cloud visualisations

In his bachelor thesis Christian Kisters [6] examined the effect of a word cloud like display variant for result surrogate. Word clouds or weighted lists display textual data by using font weight and/or size to highlight the most important or frequent terms. This allows users to

quickly find the most prominent words. Kisters evaluated a version of the result list where each item receives a similar treatment. He compared two versions of query–biased summaries (one unformatted and one using different font–sizes according to the relative frequency of terms compared to the terms of the entire result list). Highlighting of query terms was provided for both versions.

During the experiment the participants were asked to do relevance judgements for three search tasks with each of the variants. The order was rotated and the particpants were given three minutes for each combination during which they had to give correctly provide relevance judgements for as many result items as possibly based on the information shown. A total of 18 participants took part in the experiment. No significant differences in the ratio of documents correctly found relevant compared to all documents or in the ratio of correctly classified documents among the documents classified as relevant was found. However, for each of the three tasks the participants were able to classify significantly (p=.*038*) more documents with the word cloud like display variant of the summary than with the baseline (on average 24 documents vs. 20 documents). It was also the preferred display variant in terms of user experience.

6 Collaborative components and personal library

For the first evaluation of the interface components an initial, task–driven usability study was conducted with 9 participants. The main goal was to find usability and learnability problems and it used an SUS questionnaire. The study in 2012 resulted in an SUS score of 63.6.

Since then, many improvements have been made to the document collection management and sharing components, and interface components for commenting on documents and messaging other users have been added. The usability problems discovered during the first evaluation have been addressed. Therefore, for the second evaluation phase a new usability study was conducted during the month of June 2014.

This new study was combined with the usability evaluation described in Section 2. Tasks from the 2012 study were adapted to the new resources and expanded to include newly developed features. A translation of the full task descriptions is included in the appendix (see Appendix B). For a detailed description of the study participants we refer the reader to Section 2.

Task	D3.2	Error rate	D3.7	Error rate
Save doc	2	.286	2	.188
Open Plib	3	.143	3	.000
Tagging	3+4	.214	4+5+9.a	.166
Create group	6	.143	6	.094
Adding users	7+10	.214	6.a+7.c	.125
Sharing	5	.143	6.b	.000
User search	8	.428	7.a	.188
Annotating	—	—	7.b+8.b	.156
Filtering	11-13	.143	8.a+9.a+10	.109
Translation	—	—	9.b+c+f	.312
Edit Settings	14	.572	9.d+10	.344

Table 2: Error rates for usability experiment of collaboration tools. For D3.2 and D3.7 the corresponding task numbers are given (see Appendix B).

Again, the sessions were captured with screen recording software and an eyetracker. As before, an SUS questionnaire was given to the participants after all tasks were completed. The average score over all participants was 74.4, slightly higher than for the first part of the experiment (but not significantly so) and a clear improvement compared to the formative evaluation results from 2012. The score corresponds to a “good” rating according to [1].

6.1 Individual Tasks

Analysis of screen recordings, logs and observers’ notes revealed a number of interesting results, described below. The average task completion rate was .875 with a standard deviation of .066. This was in fact higher than for the first part of the experiment (described in Section 2) and suggests that the participants were quickly becoming familiar with the software.

Task specific error rates have been compiled in Table 2 and compared with the error rates from the first evaluation phase where applicable. Tasks using functionality that was only added after the first evaluations have no corresponding task in D3.2 (experiment C).

Group creation and adding users

The group tool was significantly improved and reworked in the time since the first usability evaluation. Refreshing the members list was made automatic and joining or leaving groups is now just a matter of clicking a button. Of the 20 participants of the 2014 evaluation, only one failed in creating a group, although two created a public group instead of a private group (as demanded by the task). Five users needed help with adding new users by login name. None of the participants had any trouble adding a group user from the user's profile page.

Using the personal library

Since during the previous evaluation some tasks were unclear to participants, more care was taken to explain the concept and purpose of the personal library, the pre–seeding of the library with documents for the evaluation and the difference between using the library vs. saving a document to the local hard drive. Only three users failed to save documents in the personal library. Of the successful users about half intuitively used the context menu on a result item (similar to how they would copy documents in a file manager or save a link target in a web browser). The remaining users used the save button in the document preview. All users were able to find and open the personal library from the menu, whereas a third of the user had problems finding the personal library during the previous evaluation.

In interacting with the personal library (shown in Figure 4), no user had problems deleting documents (either by using the context menu or the delete key). Only one participant failed to add tags and all participants were able to share documents with other users.

Settings and personal profiles

During the 2012 evaluation two thirds of the participants were unable to find the profile settings without help. The menu option for the settings dialog has since been moved out of the file menu and placed in the upper right hand corner labeled with the user's account name. This is similar to how profile settings are presented on many websites. In addition settings buttons were added to prominent positions for components that have additional settings (among them the profile page of a user). As a consequence, during this evaluation only 2 out of 20 users were unable to find the profile settings.

6.2 Collaboration vs. privacy

Several collaborative extensions and privacy features were developed and integrated into the browser version of the Khresmoi Professional client (see [3] for a description of the browser client). In his master's thesis Noel Kamda investigated the conflicting demands of collaboration and privacy in a health–centric search engine [5]. The support for collaboration includes an integrated user registration, including the editing of a personal user profile, a user inbox and a system for sending private messages through the system, the sharing, rating and discussion of documents, the creation of user groups and discussion forums for user groups. To ensure privacy, four concepts were implemented:

Privacy Awareness The concept of *Privacy by Default* designates all aspects of a user's profile

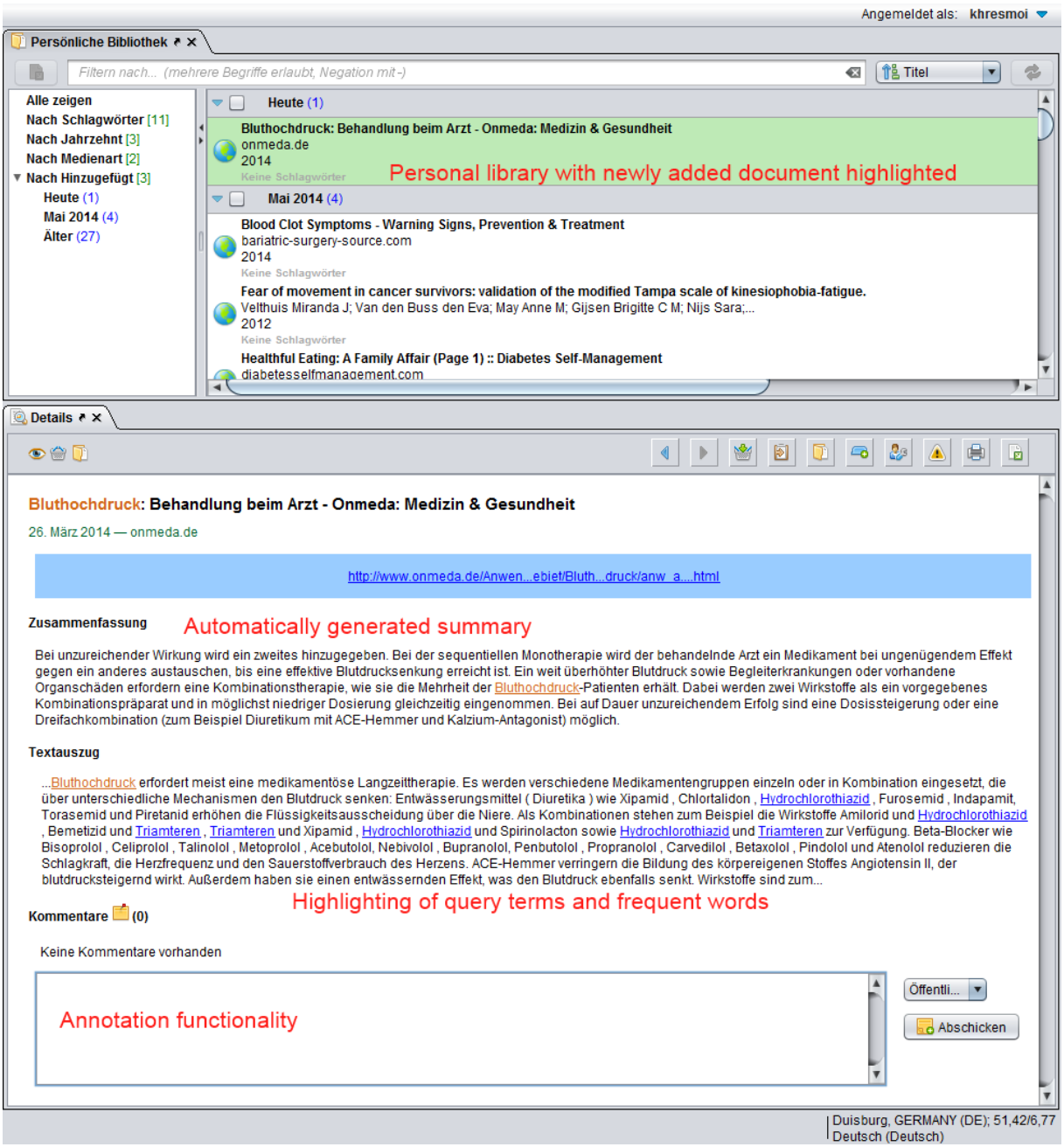

Figure 4: The right side of the Khresmoi Professional desktop client. The interface shows the personal library with the most recently added document highlighted (fades with time). The details for a document present an automatically generated summary and have query terms and frequent words highlighted and linked.

as hidden by default, so that users have to take explicit action to reveal any personal information.

Anonymity An anonymity score during profile creation (or editing) shows users how anonymous they will be based on the publicly revealed information (such as gender, country,

first name). This score is based on the total number of users with identical or similar public information without revealing the actual number.

Unlinkability The system id of the user will be encrypted on all public pages using a random algorithms, so that an attacker cannot directly link activities of the same user on multiple pages and thus create a user profile that reveals more information than the user intended.

Pseudonymity The display name of the user on public pages will be randomly pseudonymized, so that the display name cannot be used directly to link different user contributions or discussion posts.

A web–based survey was conducted with 43 participants to find out what information from a discussion post or user annotation is most often used to decide about its trustworthiness. The survey participants were asked to rate various aspects on a five point scale ranging from 1 (not at all useful) to 5 (very useful). Nearly all rated the actual content of a post with the highest possible marks (mean 4.5, mode 5), followed by user votes or user ratings on the post (mean 2.85, mode 4) and the author (mean 2.75, mode 4). It is clear that hiding or pseudonymizing user names will somewhat hinder collaboration and make it more difficult to trust such a contribution, but user generated content can probably still be useful in a system that hides user names—e.g. by allowing other users to rate the content.

The survey also found that the participants use the user search features of on-line communities mostly to find users they already know (87%), while only a minority tries to actively find new people with shared interests based on their profile (32%). For searching they often use the real name, the user name (if known) or rarely the user's email address. Asked what they are willing to reveal about themselves as part of an on-line community's profile, 46% would never divulge illnesses (40% only to friends or acquaintances), 27% would never divulge their age (57% to friends or acquaintances), 23% religion (40%), 20% occupation (83%), and 18% their marital or relationship status (59%). On the other hand, most respondents were fine with revealing their gender or country of residence on their profile. 51% were not comfortable with showing their first name for anyone but their friends within the community, and 60% said the same about their last name.

6.3 Summary

A second evaluation of the collaborative interface components was conducted and showed improvement of the SUS score from 63.6 to 74.4 points (a rating of "good") compared to the previous evaluation, likely a result of many usability changes made in the time between the experiments. Task specific error rates were generally lower compared to the rates from the first round of user evaluations. The personal library and the group tool had been reworked extensively and showed improvements, enabling almost all users to carry out basic collaborative actions without problems. Collaborative components have also been tested in other clients, and the conflict between enabling collaboration and retaining control of personal information has been investigated by Kamda [5].

7 Conclusion

This deliverable details the findings from the second user–centred evaluation in WP3. Where applicable, experiments were designed to yield results comparable to the fist round of user evaluations. Overall, the experiments showed a better user experience and higher task completion rates. A summary of the findings is given in Table 3.

When designing an interface with comprehensive functionality beyond todays standard web search, users will need some time to get used to this kind of system. Not all features may be useful to all people and not everyone may discover and comprehend features equally fast—often a problem in usability experiments with limited time for exploration. A comprehensive interface is not easy to use at first sight, but it is learnable (high learnability score on SUS). Most users were able to deduce all key elements of the interface without additional hints. While a comprehensive interface may be difficult to completely understand at first sight, users of a highly specialised system—such as Khresmoi Professional—can benefit from the added functionality (making the time investment for learning to use the system worth it).

Evaluation	Tasks	N	Main outcomes
Usability evaluation of components for search specification and result manipulation	3.2	20	*user satisfaction:* SUS 72.1 *effectiveness:* Task completion rate 75.2%
Evaluation of effect of word cloud visualization	3.2	18	*user satisfaction:* users preferred word cloud *effectiveness:* no significant differences *efficiency:* Documents per session 24 vs. 20
Evaluation of proactive search suggestion components	3.2	18	*user satisfaction:* SUS score 86.6 – 91 *effectiveness (translation):* Session recall improved by .053 Session precision improved by .08
Evaluation of combined strategic and tactical search support components	3.2	22	*user satisfaction:* 81% of users found assistance helpful for current and useful for future searches *effectiveness:* Task completion rate 95% vs. 54%
Usability evaluation of components for collection management and collaboration	3.4	20	*user satisfaction:* SUS score 74.4 *effectiveness:* Task completion rate 87.5%
Comparative evaluation of components for updating translations	3.3/4	16	*user satisfaction:* users preferred first display variant *effectiveness/effectiveness:* no significant differences

Table 3: Summary table of user experiments conducted

References

[1] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[2] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Lorraine Goeuriot, Jessica Ignalski, Matthias Jordan, Liadh Kelly, and Sascha Kriewel. Report on results of the WP3 first evaluation phase. Khresmoi Deliverable D3.2, 8 2012.

[3] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. Final flexible user interface framework and documentation. Khresmoi Deliverable D3.6, 2 2014.

[4] Jaroslaw Gustak. Proaktive Such-Unterstützung in ezDL. Diplomarbeit, University of Duisburg-Essen, Information Engineering, 2013.

[5] Noel Kamda. Collaboration and privacy in a health-centric search system. Master's thesis, University of Duisburg-Essen, Information Engineering, 2014.

[6] Christian Kisters. Implementation und Evaluierung von Dokument-Surrogaten mit Word-Cloud-Darstellung. Bachelor's thesis, University of Duisburg-Essen, Information Engineering, 2013.

[7] Sascha Kriewel. *Unterstützung beim Finden und Durchführen von Suchstrategien in Digitalen Bibliotheken*. PhD thesis, University of Duisburg-Essen, 2010.

[8] Sascha Kriewel. Interaktives Retrieval und situationsabhängige Vorschläge. *Datenbank-Spektrum*, 11, 2011.

[9] Sascha Kriewel and Norbert Fuhr. An evaluation of an adaptive search suggestion system. In *32nd European Conference on Information Retrieval Research (ECIR 2010)*, pages 544–555. Springer, 2010.

[10] James R Lewis and Jeff Sauro. The factor structure of the system usability scale. In *Human Centered Design*, pages 94–103. Springer, 2009.

[11] Jeff Sauro. Confidence interval calculator for a completion rate, 2005.

[12] Jeff Sauro. Measuring usability with the system usability scale (SUS), 2011.

[13] Jeff Sauro. What is a good task-completion rate?, 2011.

[14] Andreas Tacke. Integration von strategischer Suchunterstützung auf Makro- und Mikroebene. Diplomarbeit, University of Duisburg-Essen, Information Engineering, 2013.

[15] Andreas Tacke and Sascha Kriewel. Strategische Suchunterstützung auf Makro- und Mikroebene. In *Proceedings of the IR Workshop at Lernen, Wissen, Adaption (LWA 2013)*, Oktober 2013.

[16] Andreas Tacke and Sascha Kriewel. Strategic search support on macro and micro level. *Datenbank-Spektrum*, 14(1), 2014. http://link.springer.com/article/10.1007/s13222-014-0147-0.

A Task descriptions for usability experiment (Section 2)

During the experiment the task descriptions were given in German. These are English translations.

1. Use the program to search for `Bluthochdruck Reninhemmer`.
   (a) Find the most recent publication for this topic within the results. While doing so, explain your method to the facilitator.
   (b) Some of the results are written in English. Find those English results written in easily understandable English. While doing so, explain your method to the facilitator.
2. View some of the result previews.
   (a) Go back to the second to last preview seen.
   (b) Open the original webpage in a web browser.
   (c) Return to the program.
   (d) Look at more result previews and try to determine the purpose of the coloured links (without clicking). Explain to the facilitator what you think they do and what the different link colors stand for.
3. Filter the current result list and restrict the result to publications about `Remikiren`.
4. Remove the filter and filter restrict the result to publications containing both `Aliskiren` as well as `implication`.
5. Use the program to search for pages on `Hypertonie` within `Wikipedia`. Give a definition of “Hypertonie” to the facilitator (based on what you found).
6. Use the program to search for `Hypertension`.
   (a) Switch to the image results.
   (b) Select an image and search for similar images to that one related to the current query (for `Hypertonie`).
   (c) Switch the grid–like result view back to a list view.
7. Find the list of previous search queries and return to the results for your first search.
   (a) Using filters restrict the search results to German pages from the last three years that deal with medical education. (This task had to be removed, as a software bug made completion impossible.)
   (b) Remove all restrictions.
8. Use the program to search for definitions for `RAAS`. Look for particularly trustworthy definitions, and explain your method to the facilitator.

B Task descriptions for usability experiment (Section 6)

During the experiment the task descriptions were given in German. These are English translations.

1. Use the program to search for `Blood Clot`.
2. Save two documents from the result list in your personal library.
3. If you haven't already opened the library, open it now. It already contains some documents for this experiment.
4. You can assign tags to documents. Assign the tag `Blutgerinnsel` to the two documents you've previously saved.
5. Remove the tag again.
6. Create a private user group to share documents with other users.
   (a) Add the users `gp01` and `gp02` to the group.
   (b) Share a document from your library with the group.
7. Use the program to search for the user `Sascha Kriewel`.
   (a) Look at the profile for that user.
   (b) Leave a private message for that user.
   (c) Add the user to the previously created group.
8. Find a document created by `S. Hurme` in your library.
   (a) View the document details and print it.
   (b) Add a public annotation to the document.
9. Find a document with the title `Hypertension Types - Hypertension Center - Everyday Health` in your library.
   (a) Assign the existing tag `Bluthochdruck` for the item.
   (b) Request a translation of the summary (not for the complete document).
   (c) Edit the translation for the second sentence and close the dialog.
   (d) Change the display of the translations in the settings.
   (e) Request the summary translation again. It is now displayed using a different version of the dialog.
   (f) Edit the translation for the last sentence and close the dialog.
10. Delete all documents with the tag `Bluthochdruck`.
11. Edit your user profile to add a description.