Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Meta-analysis of the second phase of empirical and user-centered evaluations

Deliverable number	*D7.3*
Dissemination level	*Public*
Delivery date	*August 2014*
Status	*Final*
Author(s)	*Lorraine Goeuriot, Allan Hanbury, Brendan Hegarty, John Hodmon, Liadh Kelly, Sascha Kriewel, Mihai Lupu, Dimitris Markonis, Pavel Pecina, Priscille Schneller*

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.

Abstract

In this deliverable we provide a meta-analysis of the evaluations that have been conducted in year 4 of the Khresmoi project, based on the updated evaluation strategy outlined in Deliverable D7.1.2. This analysis encompasses both computational and interactive evaluations, as well as system-level and component-level evaluations. The component-level evaluations include biomedical text mining and search, biomedical image mining and search, multilingual resources and information delivery, and the user interface and search specification system. The system-level evaluations include the global computational evaluation, and interactive evaluations of the three prototypes: Khresmoi for Everyone, Khresmoi Professional and Khresmoi Radiology. We focus on each of the evaluations in turn, providing key details on the evaluations including the quality of the evaluations, results obtained, etc. This leads to a discussion and overview of all evaluations. Overall we find that high quality evaluations and research have been conducted in the Khresmoi project. Some potential improvements on the evaluation resources beyond Khresmoi are proposed, as is a higher level of automation of the evaluations.

Table of Contents

List of Abbreviations

CBIR	Content-Based Image Retrieval
CISMeF	Catalogue and Index of French-speaking Health Resources
CLEF	Conference and Labs of the Evaluation Forum
CT	Computed Tomography
D*x.x*	Khresmoi deliverable number *x.x*
ezDL	Easy Access to Digital Libraries (software)
HON	Health On the Net Foundation
HES-SO	University of Applied Sciences Western Switzerland
IAA	Inter Annotator Agreement
IE	Information Extraction
IR	Information Retrieval
LM	Language Model
MAP	Mean Average Precision
MARVIN	Multi-Agent Retrieval Vagabond on Information Networks, the HON Web Crawler.
MeSH	Medical Subject Headings
Mimir	Multiparadigm Indexing and Retrieval
MT	Machine Translation
MUW	Medical University of Vienna
NA	Not Applicable
P@*n*	Precision measured at *n* returned documents
PACS	Picture Archiving and Communication System
QUIS	Questionnaire for User Interaction Satisfaction
RF	Relevance Feedback
SCA	Service Component Architecture
SOA	Service Oriented Architecture
SotA	State-of-the-Art
SUS	System Usability Scale
WRAPIN	Worldwide online Reliable Advice to Patients and Individuals (a past EU project)

1 Executive Summary

The Khresmoi system is a complex domain-specific search system for the medical domain. Within this domain, it caters to three groups of end users through three different faces:

- *Khresmoi for Everyone*: the most straightforward interface, designed for use by the general public.
- *Khresmoi Professional*: a more comprehensive interface for use by medical professionals
- *Khresmoi Radiology*: the system emphasizing visual search with radiologists being the main user group.

The creation of such an integrated domain-specific search system is a complex task requiring modelling of the domain and its users [3], as well as a specification of the system components required and their interactions [4]. The evaluation of the performance of such a system is challenging, as it involves evaluation of multiple aspects:

- *Computational component-level* evaluations are computational evaluations of the system components taken in isolation;
- *Interactive component-level* evaluations involve an evaluation of components of the user interface and their back-end by end users;
- *Computational system-level* evaluations measure the performance of the full integrated system using a computational approach;
- *Interactive system-level* evaluation involves evaluating the full system by getting end users to perform search tasks on the system in a laboratory-type setting;
- *Observational system-level* evaluation involves evaluating the full system used in an uncontrolled way within the environment for which it was designed.

In Khresmoi, an evaluation of the system from the point of view of the first four aspects is carried out. As a search system is being evaluated, the *performance* is made up of many facets, including: retrieval performance, user satisfaction and efficiency.

A distinguishing characteristic of the Khresmoi project is its implementation of a global coordinated evaluation strategy. An independent evaluation strategy was created near the beginning of the project in D7.1.1 [1], which gave recommendations on the evaluations to be carried out in the individual work packages. A meta-analysis of the first round of evaluations was done in D7.2 [14], in which the reported results of the evaluations performed were located in the Khresmoi deliverables and compared to the recommendations in the evaluation strategy. Based on the results of the meta-analysis, an updated evaluation strategy including approaches to solve the shortcomings was presented in D7.1.2 [15].

Computational component-level evaluations can be divided into three categories, specifically: biomedical text mining and search, biomedical image mining and search, and multilingual resources and information delivery. *Interactive component-level evaluations* assessed the components of the adaptive user interface allowing search refinement, update of translations, collaborative search and effective result presentation. All of these evaluations were well carried out and produced results at the state-of-the-art or beyond. The main shortcoming of these evaluations was the lack of completely relevant evaluation resources for the text mining and search evaluations. It is important to invest more resources beyond the Khresmoi project in the creation of reliable, reusable test data.

The purpose of the *computational system-level evaluations* is to examine the backend Khresmoi system and the impact of different components on the retrieval process – from user-entered query to retrieved results. In so doing we look at the information retrieval component, the query spell corrector component and the query translation component. For the 2D radiology image retrieval evaluation, the result that introducing spelling errors has different impact in the four languages tested (English,

French, German, Czech) was obtained. Furthermore, it was found that automatically correcting the spelling has a positive impact on all four languages, but a different magnitude of impact for each language.

For the *interactive system-level evaluation*, the three faces of the Khresmoi system were evaluated by end users belonging to the target groups: members of the general public, physicians and radiologists. Typical user tasks were generated and performed by the end users while their interactions with the system were recorded and feedback was obtained through questionnaires. The main criticism of the last round of these evaluations was the low number of end users participating, so an emphasis was placed on recruiting sufficient participants in year 4. For the general public, the year 4 evaluations had 63 participants, compared to 28 participants for the year 2 evaluations. 84 physicians participated in the year 4 evaluations, compared to 14 participants in the year 2 evaluations. Finally, 26 radiologists participated in the year 4 evaluations, compared to 17 radiologists in the year 2 evaluations. The large increase in participation in the physician use case is due to the new approach developed and adopted after the difficulties faced in recruiting participants in the year 2 evaluations. Conducting the evaluations at physician symposia proved to be an effective way to attract participation from the target group, however the increased participation had to be balanced by a maximum evaluation session length of 20 minutes at such an event.

The adopted approach of examining the performance at both component-level and system-level, as well as with interactive and computational experiments, gave useful insights into optimising the overall system design. It was noted, for example, that even though the spell checkers often seem to damage the queries themselves, their use to correct misspelled queries submitted to a search engine does lead to more relevant results. Similarly, input from physicians participating in interactive evaluations led to the identification of additional resources to index and hence improved the ability of the system to provide relevant results to certain queries.

The approach adopted to implement this coordinated evaluation strategy was heavily manual, but given that this is the first time that such a comprehensive evaluation strategy was implemented for a domain-specific search system, this could not be avoided. The participants gained valuable knowledge and experience in doing this, which was particularly noticeable in improvements in the procedure between the first and second rounds of meta-analysis. The fact that the first meta-analysis identified key shortcomings that were overcome during the second round of evaluations, in particular in a very creative way for the interactive evaluation of Khresmoi Professional, supports its usefulness.

A promising way forward is to automate more of the evaluation process. Given an architecture for component integration such as the one adopted in Khresmoi, it would be interesting to run regular automated evaluations of the components and of the combinations of components, similar to the component-level approach adopted in [16]. This could be combined with an approach in which end users have access to the prototypes and can carry out searches on them, an approach currently promoted by the Living Labs initiative [17]. Work toward the end of the Khresmoi project on attracting medical professionals to participate in online evaluations and creating networks of these professionals through social media demonstrated the feasibility of Living Labs, even for busy professionals such as physicians.

2 Introduction

Khresmoi is a large scale project concerned with the development of a multilingual and multimodal medical search system which encompasses many components such as knowledge bases storing relations between medical terms, document indexes, translation components, user interfaces to access medical documents, medical image search, etc. Individual components are brought together in a private cloud and used to assist three target use case groups with their medical information needs. This

section provides a general introduction to the evaluation of the Khresmoi system, with parts adapted from D7.2 [14].

The Khresmoi system is a complex domain-specific search system for the medical domain. Within this domain, it caters to three groups of end users through three different faces:

- *Khresmoi for Everyone*: the most straightforward interface, designed for use by the general public.
- *Khresmoi Professional*: a more comprehensive interface for use by medical professionals
- *Khresmoi Radiology*: the system emphasizing visual search with radiologists being the main user group.

The creation of such an integrated domain-specific search system is a complex task requiring modelling of the domain and its users [3], as well as a specification of the system components required and their interactions [4]. The evaluation of the performance of such a system is challenging, as it involves evaluation of the aspects illustrated in Figure 1. As a search system is being evaluated, the *performance* is made up of many facets, including: retrieval performance, user satisfaction and efficiency.

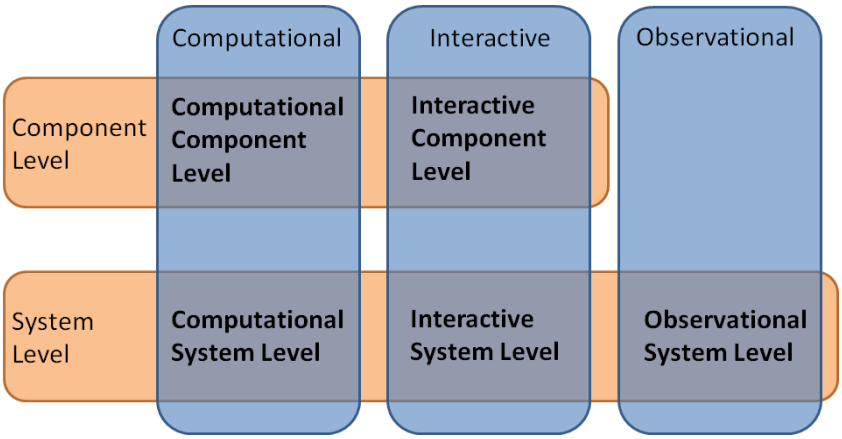


Figure 1. The aspects of the evaluation of an integrated search system.

The various aspects illustrated in Figure are now explained. *Component-level* evaluations are done at the level of the individual components making up the system (e.g. machine translation component, search component), independently of any other components of the system. *System-level* evaluations, on the other hand, consider the performance of the fully integrated system, and hence also the interactions between the components of the system. It is well known in the Information Retrieval area that simply integrating well-performing components is not sufficient to guarantee a well-performing system, due to the component interaction aspect of the system [5].

The evaluations can be done as:

- *Computational* evaluations in which performance metrics are measured using existing ground truth without considering human interaction with the system (except to potentially create the ground truth);
- *Interactive* evaluations in which the interaction between the user and system in a controlled environment is the focus[1];

[1] Note that the nomenclature for these two aspects has changed since D7.2 due to various discussions on the suitability of the "empirical" and "user-centred" labels used in D7.2 for what they should be describing.

- *Observational* evaluations in which the experiments are carried out by observing the users in an uncontrolled, "natural" environment.

These three evaluation types are a clustered version of the research continuum for interactive information retrieval research presented by Kelly [11, page 10].

Finally, at the intersections of the horizontal and vertical bars in Figure :

- *Computational component-level* evaluations are computational evaluations of the system components taken in isolation;
- *Interactive component-level* evaluations involve an evaluation of components of the user interface and their back-end by end users;
- *Computational system-level* evaluations measure the performance of the full integrated system using a computational approach;
- *Interactive system-level* evaluation involves evaluating the full system by getting end users to perform search tasks on the system in a laboratory-type setting;
- *Observational system-level* evaluation involves evaluating the full system used in an uncontrolled way within the environment for which it was designed.

An observational evaluation at component level does not make sense, as components of a search system would not be used individually in such a setting, hence there is no intersection for these aspects in the diagram.

In Khresmoi, an evaluation of the system is carried out from the point of view of the four combinations of the computational and interactive types and the system and component levels. Computational evaluations are carried out on various components of the system, and system-level computational evaluations examine how the components affect each other for both text search and image/multimodal search. Extensive system-level interactive evaluations of all three prototypes carried out with the actual targeted end users of the prototypes, i.e. members of the general public and patients, medical doctors in general, and radiologists as the representatives of the speciality targeted by one of the prototypes. An emphasis was put on using fully qualified representatives from the medical domain in these experiments, not surrogates in the form of e.g. medical students. Interactive evaluations were also carried out for components of the user interface, but these were done with students, as no specialised knowledge was required. Observational experiments are not done in Khresmoi – the experiments with users are all carried out in controlled environments as the prototypes are not currently used routinely by the end users. However, it is interesting to note the shift toward a less controlled setting in the year 4 interactive evaluations compared to the year 2 evaluations, as participants were given some free unconstrained tasks in year 4.

3 Methodology

A distinguishing characteristic of the Khresmoi project is its implementation of a global coordinated evaluation strategy. An independent evaluation strategy was created near the beginning of the project in D7.1.1 [1], which gave recommendations on the evaluations to be carried out in the individual work packages. A meta-analysis of the first round of evaluations was done in D7.2 [14], in which the reported results of the evaluations performed were located in the Khresmoi deliverables and compared to the recommendations in the evaluation strategy. This meta-analysis identified four main shortcomings in the first round of evaluations:

1. Unavailability of suitable or sufficient test data for some cases;
2. Omission of the calculation of metrics for some cases;
3. Some components not performing as well as required;
4. An insufficient number of end users in the first round of interactive system-level evaluations.

Based on the results of the meta-analysis, an updated evaluation strategy including approaches to solve the shortcomings was presented in D7.1.2 [15]. D7.1.2 proposes the following groups of evaluations:

- Component-level
  - Biomedical text mining and search
  - Biomedical image mining and search
  - Multilingual resources and information delivery
  - User interface and search specification system
- System-level
  - Global computational evaluation
  - Interactive evaluation of Khresmoi for Everyone
  - Interactive evaluation of Khresmoi Professional
  - Interactive evaluation of Khresmoi Radiology

This document provides a meta-analysis of all evaluations carried out in the second round of evaluations. To compile this document, the following procedure was used:

1. Groups performing evaluations were reminded of the updated evaluation strategy (D7.1.2) at a plenary project meeting.
2. An information gathering form (Figure) was sent to each group performing an evaluation defined in D7.1.2. This form served as a reminder of the requirements of the evaluation to carry out and as a means to gather the results in a structured way.
3. A person independent of those performing the evaluation was assigned to each group of evaluations. This person collected the completed forms, analysed them, verified the results and compiled the results into the template shown in Figure 1.
4. The deliverable coordinator created global summaries of all evaluations.

Beyond the meta-analysis, the document also presents new results of system-level computational evaluations of the prototypes.

The remainder of this document is structured as follows. Section 4 presents a meta-analysis of the component-level evaluations, with the computational evaluations covered in Section 4.1, and the interactive evaluations covered in Section 4.2. Section 4.3 summarises the component-level evaluations. Section 5 presents the system-level evaluations in an analogous way. Section 6 concludes the document.

Evaluation:	<evaluation name to go here -- ; separate form should be completed for each evaluation being completed in the WP>					
Deliverable(s) and sections						
Criteria ID	**Title**	**Strategy - as described in D7.1.2**	**Deliverable**	**Section**	**Evidence (in support of what should have been included in the evaluation, as described in Column C's strategy)**	**Follow-up (what is missing)**
1	Evaluation Points					
2	Related tasks					
3	Evaluation Style					
4	Exisiting resources					
5	Exisiting evaluation campaigns					
6	Resources created/to be created by Khresmoi					
7	Evaluation measures					
8	Recommendations					
9	Link to objective measurement of the progress toward project objectives					
10	Details on evaluation to be conducted					
Evaluation Criteria						
11	Conclusions from deliverable					
12	Baseline					
13	Start of the Art					
14	Relevant to Khresmoi					
15	Summary					
16	**Is there anyhing missing?**					

Figure 2. Information gathering form for each evaluation.

Evaluation points: Evaluation points in the diagrams on pages 8–10 of D7.1.2 (reproduced in the appendix of this document for ease of reference).

Khresmoi tasks: The Khresmoi tasks to which the evaluation corresponds.

Relevant deliverables and sections: The Khresmoi deliverable sections in which the descriptions and results of the evaluations are to be found

Resources used/created: Resources used for the evaluation, pointing out if publicly available resources were used, or if resources for the evaluation were created in Khresmoi.

Description of experiment(s): A summary of the experiments carried out for the evaluation

Evaluation measures: The evaluation measures used for the evaluation

Results: A summary of the results obtained from the experiments

Conclusion: A short conclusion on the results of the evaluation

Recommendations: These recommendations indicate potential lines of investigation for projects after Khresmoi.

Figure 3. Summary template used in this document for each analysed evaluation.

4 Meta-Analysis of Component-Level Evaluations

The component-level evaluation evaluates the components of the Khresmoi system in isolation. In this section we provide details on both the empirical and user-centred component-level evaluations conducted in the second round of evaluations in the Khresmoi project.

4.1 Computational Component-Level Evaluations

These evaluations were defined in Khresmoi deliverable, D7.1.2, updating D7.1.1 and including the outcome of the first meta-analysis in D7.2. These evaluations can be divided into three categories, specifically: biomedical text mining and search, biomedical image mining and search, and multilingual resources and information delivery. In this section we review these evaluations. The evaluation titles below correspond to the planned evaluations described in D7.1.2.

The knowledge base performance evaluation did not form part of the evaluation strategy (in D7.1.2). Naturally the knowledge base was also evaluated for performance, speed, reliability, etc. and a summary is also included in this section.

4.1.1 Biomedical Text Mining and Search

The objective of this section is to summarize and analyze evaluation efforts for text processing tools.

4.1.1.1 Semantic annotation of literature

Evaluation points: E3.4

Khresmoi tasks: T1.1, T1.2

Relevant deliverables and sections: D1.8

Resources used/created: Manually corrected annotations have been created (See Section 2.2 of D1.8): "The Khresmoi annotation pipeline was developed in an iterative manner. At each development iteration, annotations created by the pipeline were corrected by human annotators. These corrections were used to drive further development iterations. By comparing the automatic annotations with the corrections at each iteration, we can provide a measure of system performance." One observation here is that the way manual annotations are made is not optimal, as observed in the previous version of this deliverable, but that is compensated by the comparison with the CALBC corpora.

Description of experiment(s): The Khresmoi annotation has been compared with gold and silver standards in order to obtain effectiveness measures.

Evaluation measures: Precision, Recall, F-measure, inter-annotator agreement (IAA).

Results: Results are compared with existing state-of-the-art. Comparison not necessarily advantageous, but justifiable by the fact that Khresmoi, unlike the referenced state-of-the-art, did not optimize for the CALBC test collection

Conclusion: High level of agreement between manual and automatic annotations. When comparing with an external resource (CALBC Silver Standard), Khresmoi compares well, showing similar performance as other systems. Additional error analysis indicates that the disagreement between Khresmoi and CALBC annotations are largely due to different understandings of what is to be annotated.

Evaluation of the entity annotation has been improved. While it was not tested with gold standard (there was none to also contain CUI annotations) it was compared with a silver standard created by other systems and with a manual correction created in-house

Recommendations: Significant effort is required for creating gold standards, perhaps in the context of a targeted evaluation campaign.

4.1.1.2 Textual search & ranking

Evaluation points: E1.3

Khresmoi tasks: T1.4

Relevant deliverables and sections: D1.8

Resources used/created: Query logs extracted from the online system: "The dataset used for this experiment is extracted from recorded ezDL user logs. We utilize the users' queries and their search behaviour to examine the effectiveness of the integrated model."

Description of experiment(s): Retrieval experiments were done based on user logs and evaluated considering clicked documents to be relevant. The baseline was the existing Khresmoi system. The experiments did not use any of the existing test collections because none of them contained user information, essential to the collaborative filtering method applied here.

Evaluation measures: Mean Average Precision (MAP), P@5,P@10,P@20

Results: A new retrieval method is proposed here and the evaluation shows it is able to improve against a baseline consisting of the existing method. The evaluation collection is created from logs - which is problematic in itself, but a good initial effort.

Conclusion: The new evaluation presented in D1.8 for the problem of search is limited in its comparison with state of the art in the domain because user data is considered - which is generally not available on public evaluation campaigns. The state of the art is either on the collaborative filtering or on medical retrieval, and therefore there is no direct comparison we can make here. However, this is

very relevant to Khresmoi because ultimately it will have access to large amounts of user logs from which to improve its retrieval method

Recommendations: It would be good to validate that the qrel used here is indeed an indicator of relevance. Typically, click data is a mix of relevant and non-relevant information.

4.1.1.3 Document categorizer

Evaluation points: E3.5

Khresmoi tasks: T1.5

Relevant deliverables and sections: D1.8, D1.6

Resources used/created: HON data regarding the trustability of websites for French, Spanish, German, Italian and Dutch

Description of experiment(s): The evaluation of trustability classification has been extended to other languages. Performance is maintained where enough data is available, but degrades with the reduction in resources

Evaluation measures: Precision, Recall, F1

Results: Comparable with English corpora, provided enough data is available

Conclusion: The new evaluation follows what has already been done for English, and presented in D1.6. It now includes 5 additional languages. There is no discussion of the state of the art, but the results seem very promising if there is enough data.

Recommendations: The issue of trustability is very important but very difficult to assess. The current experiments are a good start in this direction, but further work is necessary to create a repeatable benchmark.

4.1.2 Biomedical Image Mining and Search

The objective of this section is to summarize and analyze the evaluation efforts for image mining and analysis tools.

4.1.2.1 Evaluation of learning from relevance feedback

Evaluation points: E2.10

Khresmoi tasks: T2.5

Relevant deliverables and sections: D2.7

Resources used/created:

- ImageCLEF 2012 medical image retrieval
- ImageCLEF 2008-2011 medical image retrieval (http://www.imageclef.org/2008/medical)
- AMIA ImageCLEF 2013 medical image retrieval (http://www.imageclef.org/2013/medical)

Description of experiment(s): Empirical evaluations of multimodal image retrieval tools using relevance feedback have been performed. Results were compared against a baseline consisting of self-made text retrieval based on Lucene as well as existing runs on the used collection (ImageCLEF 2012 medical image retrieval)

Evaluation measures: Mean Average Precision

Results: MAP up to 0.31 was achieved by the mixed runs (using both visual and text features) using relevance relevance feedback and late fusion. This is a significant improvement over using image features alone and a minor improvement over text features.

Conclusion: Relevance Feedback **(**RF) has been shown to work well for medical image search tasks. Even when using visual features only, despite still not going beyond text-based methods, RF has improved effectiveness.

Recommendations: Further work is needed in fusing data modalities. It is already shown that multi-modal search can outperform text-only search, but there is still a large space to investigate in how to optimally merge modalities.

4.1.2.2 Evaluation of classification of radiology images

Evaluation points: E3.9, E3.10

Khresmoi tasks: T9.2, T9.3

Relevant deliverables and sections: D2.7

Resources used/created:

Lung Tissue Research Consortium (LTRC) dataset.[1]

Additionally, the following resource were created:

- A reference multimedia database for interstitial lung diseases
- High resolution lung CT data set (MUW) with case level annotations of pathology patterns

Description of experiment(s): Image classification experiments were performed and compared to various baselines: (1) prior results on same data set (LTRC [2]) (2) 2-layer convolution neuralLTRC: network.

Evaluation measures: voxel-wise misclassification error, image-level misclassification error

Results: The new system improved classification performance over the state of the art, inclusion of large-scale non-annotated data improves accuracy on test set. Over-segmentation improves results for classification (super-voxel).

The image based tissue classification in an exemplary case of lung diseases could be improved beyond the state of the art in the project. It demonstrates that the overall approach of leveraging large amounts of non-annotated data to improve the analysis, retrieval, and classification of imaging data causes a substantial improvement.

Conclusion: (1) Adding layers to a CNN significantly improves performance in terms of misclassification error. Overall achieved misclassification error of approximately 5% (2) Unsupervised pre-training of networks improves accuracy even if conducted across sites.

Recommendations: A future system would greatly benefit from the combination of the location and classification work.

4.1.2.3 Semantic annotation of images

Evaluation points: E3.6

Khresmoi tasks: T1.1, T1.2, T2.4

Relevant deliverables and sections: D2.1, D2.7

[1] http://www.insight-journal.org/browse/publication/109

Resources used/created: full body CTs with annotations, provided by the Visceral project.[1]

In D2.1 additional resources had been created: Clinical CTs with annotated locations of image centers (n~4000), annotated landmarks on a set of full-body CTs.

Description of experiment(s): The semantic annotation of images is here a segment labelling task. The evaluation is done on the only existing public resource available and the results show comparable results with other runs using the same collection.

Evaluation measures: Dice coefficient, location error (distance)

Results: The results obtained were comparable with the state of the art described in the context of the VISCERAL evaluation campaign[2].

Conclusion: Localization of anatomical structures in medical imaging data was improved both on a coarse detection level, and on a fine localization level. This is a necessary prerequisite for organ specific classification and retrieval [8, 9, 10]. The labeling accuracy is sufficient for a coarse segmentation of the imaging data, and is necessary for retrieval.

Recommendations: Further evaluation metrics could be considered, to increase comparability with VISCERAL runs.

4.1.3 Multilingual Resources and Information Delivery

This section provides an overview of the two evaluations conducted in the multilingual resources and information delivery group.

4.1.3.1 Summarization

No experiment was done.

4.1.3.2 Evaluating summary translation

Evaluation points: E1.5, E1.9, E2.6, E2.8

Khresmoi tasks: T4.6

Relevant deliverables and sections: D4.7

Resources used/created: 1,500 sentences from summaries generated with Khresmoi summarizer have been manually translated in Czech, French and German. The summaries were generated from CLEF eHealth 2013 task 3 dataset [12], query and document ID were provided with the sentences. This dataset has been used for a shared task in a workshop on Statistical Machine Translation [13].

Description of experiment(s): CUNI provided a baseline for the WMT task, as they were the organizers, and also performed some internal evaluations with the dataset. The experiments consist in training a statistical MT system on a training subset of the WMT dataset, and evaluating it on the remaining subset. Details are given in D4.7.

Evaluation measures: BLEU, TER, PER and CDER (automatic evaluation metrics)

Results: CUNI provided a strong baseline for the shared task, and got some improvement over the baselines afterwards, by training additional models and adding them to the interpolation.

[1] http://visceral.eu
[2] http://www.visceral.eu/assets/assets/VISCERAL-at-ISBI-segmentation-results.pdf

Conclusion: The development of this evaluation task and the resulting experiments allowed CUNI to improve the Khresmoi MT system and adapt it to the translation of summaries. The results obtained are comparable to the state-of-the-art results.

Recommendations: A manual evaluation of the translation of summaries would be useful, to assess whether Khresmoi users can read and understand them.

4.1.3.3 Spelling correction

Evaluation points: E1.05, E2.05

Khresmoi tasks: T8.3

Relevant deliverables and sections: Part of the global computational evaluation (Section 5.1.2)

Resources used/created: A test set of misspelled English queries has been automatically generated from HON and TRIP query logs (50 patient short queries, 50 patient long queries, 50 physician queries). Each query comes with several misspelled versions with varying degradation levels.

A domain-specific dictionary based on the eHealth 2013 collection has been generated from a Lucene index of this collection.

Description of experiment(s): Khresmoi spellchecker is run on the misspelled English queries, as well as other state-of-the-art spellcheckers for comparison purposes. The spellcheckers Khresmoi is compared to are:

- Lucene Spellchecker – the default Lucene Spellchecker, using the dictionary created on the eHealth 2013 collection
- HunspellChecker – the state-of-the-art spellchecker used in many open source projects (e.g. OpenOffice, Mozilla)

The output of the spellcheckers is compared to the initial queries by:

- **edit distance:** between the original query and the spell-corrected query.
- **number** of:
    - **corrected terms**: terms which were misspelled and were corrected to their original form
    - **non corrected terms**: terms which were misspelled but even after correction are still different from the original
    - **mis-corrected terms**: terms which were not misspelled, but were corrected to something else

Evaluation measures: Edit distance, #corrected terms, #non-corrected terms, #mis-corrected terms

Results: The spellcheckers actually damage almost all queries. For general practitioner queries the edit distance is best using the Khresmoi (HON) spellchecker and then the Lucene spellchecker, followed by Hunspell. For general public queries, the order of the spell checkers' results is the same, but with only a couple of exceptions, all worse than the original misspelled queries.

In terms of number of corrected terms however, Hunspell surprisingly performs second best (after Khresmoi), on average, followed by Lucene. Similar ranking for the number of miscorrected terms.

Conclusion: As expected, the dictionaries play a pivotal role in spellchecking. A domain specific dictionary is important, but often too aggressive in its correction.

Recommendations: The Khresmoi spellchecker outperforms Lucene and Hunspell, but the spelling corrections are still done on a term-by-term basis. It is probable that even better scores may be obtained if the spellcheckers would take more context into account when generating suggestions.

4.1.4 Knowledge Base

In this section we analyse the knowledge base evaluation conducted in WP5 of the Khresmoi project. The evaluation of the knowledge base did not form part of the evaluation strategy, but is included here for completeness.

4.1.4.1 Second phase indexing workflow performance evaluation

Evaluation points: NA

Khresmoi tasks: T5.1, T1.3.

Relevant deliverables and sections: D5.6, section 5.1

Resources used/created: Khresmoi medical documents crawl

Description of experiment(s): A test for the document indexing and annotation workflow was performed, comparing the indexing speed to previous processing runs.

Evaluation measures: indexing speed, documents per seconds

Results: The new version of the workflow decoupled retrieving and indexing, after fetching from CouchDB was identified as a bottleneck. Notable speed improvements were observed (speed up of factor 2-3.5)

Conclusion: The initialization of gazetteers represents an overhead that is especially noticeable for small indexing runs (it represented about 40% of the total time needed to index 31K Wikipedia documents). Without the overhead, indexing of English, German or Czech documents can be done at a speed of about 35 documents per seconds. For French or Spanish documents, where no semantic annotation is performed, this speed increases to 50 documents/seconds. This is a clear improvement on the first version of the workflow. However, the linear dependency between number of documents and time needed to process them makes the current setup unsuitable to scale up to billions of documents (solution: parallel indexing).

Recommendations: None

4.1.4.2 Second phase scalability evaluation

Evaluation points: NA

Khresmoi tasks: T5.1, T1.3.

Relevant deliverables and sections: D5.6, section 5.2

Resources used/created: Khresmoi medical documents crawl, Khresmoi knowledge base, RadioWiki dataset, Radiology reports dataset, RadLex to UMLS mappings, RadLex labels, MeSH labels, ImageClef, UMLS, Drugbank datasets

Description of experiment(s): Several services were tested for scalability regarding the number of parallel requests. A combination of simple and complex requests were issued against the services (N=5000), using 10, 100 and 1000 concurrent threads.

Evaluated services:

- SPARQL endpoint (D5.6, section 5.2.1)
- Disambiguator (D5.6, section 5.2.2)
- Quick search (D5.6, section 5.2.3)
- Co-occurrence search (D5.6, section 5.2.4)
- Semantic type-ahead search (D5.6, section 5.2.5)

Evaluation measures: average response time, response time under increasing load

Results:

SPARQL endpoint: For all query types the average response time was around 100ms, with no response time exceeding 300ms (with 1.2 billion statements in the knowledge base). Under heavy concurrent load the query response time dropped slightly.

Disambiguator: The disambiguator is used to provide suggestions while users type a query. For this the response time needs to be below 100ms to be perceived as instantaneous by users. The average measured response time was 71ms, however with increased load response times exceeded 100ms.

Quick search: The quick search service provides results to users on explicit request. Systems with response times below 1s are usually regarded as responsive. The average response time of the service was around 73ms for a single request thread (N=500). For concurrent requests the response time increases to 107ms-173ms (10-1000 users, N=5000).

Co-occurrence search: The co-occurrence search service provides results to users on explicit request. Systems with response times below 1s are usually regarded as responsive. the average response time of the service was measured as 175ms for a single request thread (N=500), for concurrent requests the response time increased to 193ms-249ms (10-1000 users, N=5000).

Semantic type-ahead search: The type-ahead service is used to provide suggestions to users while they type their query. A response time below 100ms is perceived as instantaneous in this context. The average response time of the service was measured as 85ms (N=400) for a single request thread, with 10-1000 concurrent threads the response time did not increase significantly.

Conclusion:

Significant increase for the SPARQL endpoint in average response time was only observed when going from one to 10 threads. Subsequent increases result only in minor deterioration of performance. Therefore, it can be expected that the service will scale well.

For the disambiguator response times increased significantly compared to the previous version. However, despite increased complexity of requests (to yield more precise results) the average response times for a single request thread remain below the important 100ms threshold. When scaled up to 10-1000 concurrent requests, the response time increased (above the threshold), but remained close to 100ms. Better performance could be achieved through changes in the underlying hardware infrastructure.

The other services handle large numbers of concurrent requests without severe degradation of response time. Even with 1000 concurrent request threads the response times remain well below 1sec (or 100ms for the semantic type-ahead service).

Recommendations: None

4.2 Interactive Component-Level Evaluations

The evaluations that were assessed in Work Package 3 focused on the development of the flexible adaptive user interface. The system allows users to access the Khresmoi system in different languages, provides task and user specific search assistance, whilst offering options for different result presentations for the users' needs.

This section describes the details of the evaluation that took place in the resulting components in WP3, as well as some other components developed in other work packages that are used in WP3. Evaluations are heavily user-centered, assessing evaluation measures in efficiency, effectiveness and user satisfaction. Observations, eye-tracking and logging data are available for all of the evaluations conducted, supporting the findings of the experiments.

4.2.1 Suggestion of Search Refinement and Continuation Options and Support

4.2.1.1 Evaluation of effectiveness of search refinement options

Evaluation points: E1.1, E1.10, E1.12

Khresmoi tasks: T3.2

Relevant deliverables and sections: D3.7, Section 3

Resources used/created: Khresmoi Professional medical document collection.

Description of experiment(s): 18 participants were used for this evaluation. Participants are given search tasks with predefined queries. For some tasks the users are supported by suggestion modules. Effectiveness is measured using interactive session recall and precision based on documents saved by users.

Evaluation measures: SUS score (user satisfaction), interactive session recall and precision (based on documents saved by user during session).

Results: The average SUS scores from the baseline (interactive session recall/precision without support of suggestions) for the UI evaluations was 68-69.5.

SUS scores between 86.6 and 91 were obtained using the Khresmoi search refinement options. However, interactive session recall and precision only significantly improved when using translation suggestions, but not for spelling suggestions or related queries.

Conclusion: Interaction with components was well received by participants (high usability scores on SUS), but a significant improvement of users' performance could only be shown for translation suggestion.

Recommendations: None.

4.2.1.2 Evaluation of combined support through scaffolding and tactical suggestions

Evaluation points: E1.1, E1.10, E1.12

Khresmoi tasks: T3.2

Relevant deliverables and sections: D3.7, Section 3

Resources used/created: Khresmoi for Everyone medical document collection.

Description of experiment(s): The 22 participants used in this study are given two complex search tasks. Half of the participants receive scaffolding support (both have the option to use tactical search suggestions). Performance is measured using task completion time and rate.

Evaluation measures: Task completion time, use of search features and suggestions, task completion rate, user satisfaction.

Results: Scaffolding was well received - 9 of 11 users with support found it helpful and would want to use it again. A combination of scaffolding with tactical suggestions resulted in a significantly improved task completion rate (0.95 vs. 0.54), but not in task completion time (36.36 vs. 35.73). Users were significantly more likely to use advanced search features (6.91 vs. 3.18) and tactical suggestions (2.36 vs. 1.91) when scaffolding support was provided.

Conclusion: Combining scaffolding with tactical search suggestions does not make searching faster, but can allow searchers to create more effective search strategies and provide a more pleasurable and less frustrating experience.

Recommendations: None.

4.2.2 Usability of Components for Search Specification and Result Manipulation

Evaluation points: n/a (E1.1, E1.10, E1.12 for result presentation)

Khresmoi tasks: T3.1, T3.2

Relevant deliverables and sections: D3.7, Section 2

Resources used/created: Khresmoi Professional medical document collection.

Description of experiment(s): This experiment was first conducted and reported in year 3 of the Khresmoi project and reported in D3.2. Based on recommendations provided in D7.1.2 to use a larger sample size and to include task specific error rates, the experiment was repeated in this year 4 of the Khresmoi project. A larger sample size was used for this experiment in year 4, compared to that used in year 3 of the Khresmoi project (20 vs. 8). Error rates were provided for 11 tasks this year and detailed analysis for 7 tasks provided.

It is a task driven usability study, in which users are given specific tasks to perform with the system. Usability is measured using the SUS questionnaire. Screen recording and eye-tracking are used for qualitative analysis of task performance.

Evaluation measures: SUS score (user satisfaction), task specific error rates, task completion rate, analysis of screen recordings.

Results: The baseline results are: average SUS scores over multiple tests (68 or 69.5 depending on the article consulted) and average task completion rates (78%) for UI evaluations.

This year improved SUS scores (72.1 vs. 66.9) compared to previous version of the UI were obtained. Overall ″good″ SUS score (when translated into attribute rating), split into factors: 81.7 for learnability, 69.7 for usability. Improved error rates for most tasks compared to previous version of the UI. Average task completion rate (TCR) slightly below mean (75.2%) likely due to inclusion of several complex or difficult to solve tasks.

Conclusion: Users feel that the system can be learned quickly. Overall improvement over the first interface version was found. Some aspects, such as visualisation of document attributes need further work.

Recommendations: None.

4.2.3 Usability of Components for Updating Translations

Evaluation points: E1.11, E2.12

Khresmoi tasks: T3.3, T3.4

Relevant deliverables and sections: D3.7, Section 4

Resources used/created: Khresmoi Professional medical document collection; Khresmoi MT system (en->de).

Description of experiment(s): 18 users were given translation correction tasks for two documents using two interface variants. This was a comparative evaluation to choose the most useful UI variant.

Evaluation measures: Task completion time, error rates, user satisfaction, system speed.

Results: No significant difference in errors made or task completion time, but users clearly favored parallel display of full text with highlighting for correcting/editing translations; showing all translation alternatives was criticized as information overload, users preferred to translate sentences in the context of a complete paragraph, not isolated.

Conclusion: Users clearly favored one variant, but inclusion of some features from other variant should be considered. Usability of the components is important from point of user experience and satisfaction, but less so from a point of efficiency or effectiveness, since both variants allow users to perform the necessary task and most of the time spent correcting resources is spent on the mental task, not interacting with the component.

Recommendations: None.

4.2.4 Result Presentation

4.2.4.1 Evaluation of result presentation

Evaluation points: E1.6

Khresmoi tasks: T3.2

Relevant deliverables and sections: D3.7, Section 5

Resources used/created: Khresmoi Professional medical document collection.

Description of experiment(s): This is an update on the result presentation evaluation reported in D3.2, based on recommendations in D7.2. Additional analysis of eye-tracking videos and reporting of task completion times was recommended in D7.2.

Evaluation measures: Task completion time, relevant documents identified by users, error rate.

Results: No significant differences in task completion times (221s for the grouped result list vs. 203s for the tabbed result list) were observed. Eye-tracking analysis provided no additional insights into skipping behaviour due to technical limitations.

Conclusion: No additional information was gleaned from the additional analysis conducted on the result presentation evaluation from year 3 of the Khresmoi project.

Recommendations: None.

4.2.4.2 Evaluation of word cloud visualisation

Evaluation points: E1.6

Khresmoi tasks: T3.2

Relevant deliverables and sections: D3.7, Section 5

Resources used/created: Khresmoi medical document collection.

Description of experiment(s): 18 participants took part in this experiment. Users have to identify relevant documents from result sets using different displays of query-biased summaries.

Evaluation measures: Relevant documents identified by users, error rate, user satisfaction.

Results: User performance without word cloud visualization was used as a baseline.

No significant differences found for error rate (precision) compared to the baseline, but user were able to classify significantly more documents ($p=.038$) during a fixed time period using the word cloud visualization and also preferred this display variant.

Conclusion: While the word cloud visualization does not allow more precise relevance judgements (no significant improvement on baseline), it allows users to make them faster, as the most important terms are increased in font size.

Recommendations: None.

4.2.5 Usability of Collaborative Components

Evaluation points: E1.0, E1.1, E1.2, E1.6, E1.7, E1.8, E1.10, E1.12

Khresmoi tasks: T3.2, T3.4

Relevant deliverables and sections: D3.7, Section 6

Resources used/created: Khresmoi Professional medical document collection.

Description of experiment(s): This experiment extends from that reported in D3.2 in year 3 of the Khresmoi project, based on recommendations in D7.2 to use a larger sample size than used in D3.2 (8 participants) and to include task specific error rates. In this repeat experiment 20 participants were used and error rates for 11 tasks, with detailed analysis for 3 tasks, are provided.

This is a task driven usability study (users are given specific tasks to perform with system, usability is measured using SUS questionnaire). Screen recording and eye-tracking are used for qualitative analysis of task performance.

Evaluation measures: SUS score (user satisfaction), task specific error rates, task completion rate, analysis of screen recordings.

Results: The baseline results are: average SUS scores over multiple tests (68 or 69.5 depending on the article consulted) and average task completion rates (78%) for UI evaluations.

This year improved SUS scores (74.4 vs. 63.6) compared to previous UI version were obtained; overall this was a ″good″ SUS score (when translated into attribute rating). Improved error rates for most tasks compared to previous UI version. Average Task Completion Rate of 87.5%. Use of larger sample size compared to previous study (20 vs. 9).

Conclusion: Overall improvement over the first interface version was observed.

Recommendations: None.

4.3 Summary and Analysis

In this section, we provide an overall meta-analysis of the computational and interactive component-level evaluations. Table 1 summarises all of the evaluations described in this section using a “traffic light” visualisation. Each evaluation is rated on whether the evaluation approach was satisfactory (column EV) and whether the evaluation results are satisfactory (column RE). Overall, both the evaluation approaches and the evaluation results are satisfactory for the experiments performed. Details are discussed further in this section, with the relation of the results to those of the previous round (D7.2) and the updated evaluation strategy (D7.1.2).

Results are now available for textual search and ranking experiment, as foreseen in D7.2. In the end, doing the experiments on radiology reports foreseen in D7.2 was not considered feasible due to the lack of extensive manual annotation of (German) radiology reports and insufficient resources to create them in the project, and this evaluation was removed from the updated evaluation strategy (D7.1.2). For two of the experiments in the biomedical text mining and search section, the evaluation approach has been indicated as partially satisfactory to highlight the issues remaining with the evaluation data. For the semantic annotation experiment, it is pointed out that “the way manual annotations are made is not optimal,” while for the textual search and ranking experiments, it is not clear if the assumption that a click from a search log indicates relevance is good. In both of these cases, it is important to invest more resources beyond the Khresmoi project in the creation of reliable, reusable test data.

Evaluation Category	EV	RE	EV	RE	EV	RE	EV	RE	EV	RE
Biomedical text mining and search	4.1.1.1: Semantic annotation		4.1.1.2: Textual search & ranking		4.1.1.3: Document categorizer					
Biomedical image mining and search	4.1.2.1: Relevance feedback		4.1.2.2: Image classification		4.1.2.3: Semantic annotation					
Multilingual resources and information delivery	4.1.3.1: Summarization		4.1.3.2: Summary translation		4.1.3.3: Spelling correction					
Knowledge base	4.1.4.1: Indexing workflow		4.1.4.2: Scalability							
Interactive	4.2.1: Search refinement & continuation		4.2.2: Usability of components		4.2.3: Translation update		4.2.4: Result presentation		4.2.5: Collaborative components	

Table 1. Meta-analysis of component-level evaluations. The numbers refer to sections in this document. The colours below each heading indicate the following: Was the evaluation approach satisfactory? (in column EV); Are the results satisfactory? (in column RE). Green indicates satisfactory, yellow partially satisfactory, red not satisfactory, and blue that the experiment was not performed.

For the biomedical image mining and search section, good use was made of available evaluation datasets and the results produced were state-of-the-art or beyond. Unfortunately, efficiency results were again not reported for either the text or image mining and search experiments, although the indexing and annotation efficiency is covered by the knowledge base evaluation in Section 4.1.4.1.

For the multilingual resources and information delivery section, summarization was flagged in the previous meta-analysis (D7.2) as an area where the Khresmoi-developed approach was not working optimally. However, given the resources available in the project, it was decided to adopt the alternative summarization approach with a better performance rather than further developing the Khresmoi approach. There was therefore no further evaluation of summarization done. The summary translation evaluation was well carried out through the organisation of a shared evaluation task, and led to the creation of a new evaluation data resource. The spell checker evaluation led to the recommendation to update the dictionary of the Khresmoi spell checker. In D7.2, Query Reformulation was included in this section and indicated as not done. In the updated evaluation strategy (D7.1.2), this evaluation was moved to the interactive evaluations as Search Refinement and Continuation.

Experiments on the knowledge base focussed on efficiency, in particular the indexing speed and the response time of the knowledge base for queries. All results were found to be satisfactory.

The experiments in the interactive component-level evaluations were all conducted with a sufficient number of participants, improving the main shortcoming from the first round of evaluations. Translation suggestion was shown to give a significant improvement in the user search performance, while scaffolding and tactical search suggestions do not increase the speed of obtaining results but allow searchers to use more effective search strategies and hence make searching less frustrating. The experiments on the usability of components for search specification and result manipulation and on the usability of collaborative components repeated experiments evaluating the search interface from the previous round of evaluations with more participants. An improvement over the version of the interface evaluated in the last round was found, seen through higher SUS scores. Participants generally felt that the interface can be learned rapidly.

For updating translations, participants tested two interface variants and clearly favored one variant, but potentially with the inclusion of some features from other variant. For the result presentation, the further analysis of the eye-tracking data collected in the previous round of experiments recommended in D7.2 did not yield any further insights. Word cloud visualization was found to not allow more precise relevance judgements (no significant improvement on baseline), but it allows users to make relevance judgements faster.

Translation support was indicated as an aspect of Khresmoi with insufficient evaluation in the previous meta-analysis, with the recommendation to conduct this analysis during the system-level interactive evaluations. However, with the emphasis on reducing the length of these evaluation sessions in year 4 (see Section 5.2), this was not possible, so further evaluation of the translation support components was not carried out, as indicated in the updated evaluation strategy (D7.1.2).

User profiling and personalization was indicated as not done in D7.2, and was planned in the updated evaluation strategy. However, it was not possible to do this evaluation finally: (a) due to the decision to conduct interactive evaluations of only 20 minutes with physicians, not enough information could be collected to evaluate profiling; (b) there is also a computational evaluation aspect to personalization and profiling evaluation for which ground truth needs to be collected, but there were insufficient resources to create this ground truth within the project.

5 System-Level Evaluations

In evaluating the performance of a system such as Khresmoi, it is important to gain a holistic overview of the performance of the system as a whole, both from the point of view of the user experience in using the system in a live search setting and from the viewpoint of the impact of system components on each other. In this section we provide details on both the computational and interactive system-level evaluations conducted in the last year of the Khresmoi project.

5.1 Computational System-Level Evaluations

We have conducted multiple global computational evaluations in the Khresmoi project. The purpose of these evaluations is to examine the backend Khresmoi system and the impact of different components on the retrieval process – from the user-entered query to retrieved results. In so doing we look at the information retrieval component, the image analysis component, the annotation component, the query spell corrector component and the query translation component. This section starts with an analysis of the experiments on the efficiency of the system architecture reported in a deliverable from WP6.

The final two parts of this section are different in nature to the remaining parts of this document, as they are not an analysis of experiments done by others, but experiments on the full system carried out in WP7. In year 3 of the Khresmoi project a pilot global computational evaluation was conducted for the general public and general practitioner use cases. This evaluation was reported in D7.2. In the fourth year of the Khresmoi project we conducted the full-scale version of this evaluation for the general public and general practitioner use cases. This year we also conducted a global computational evaluation for the radiology use case. The remaining two sub-sections present summaries of the experimental results of these evaluations, with full details of the experiment available in attached documents.

5.1.1 System Architecture

In this section we analyse the final evaluation conducted in the Khresmoi project on the system architecture, on the "Full Khresmoi integrated infrastructure." These evaluations were conducted in WP6 and did not form part of the overall Khresmoi evaluation strategy presented in D7.1.1.

Evaluation points: NA

Khresmoi tasks: T6.4.3, T6.4.4, T6.4.5

Relevant deliverables and sections: D6.5.3: all sections

Resources used/created: KHRESMOI Full Cloud VMs and nodes, SCA Components and services

Description of experiment(s): This evaluation is relevant for both textual search and semantic search scenarios. The method used for D6.5.2 was used as the baseline. The CPU usage was measured as CPU resource consumption during evaluation, the memory usage as memory resource consumption during evaluation, and the average response time: average response time for queries during evaluation

Evaluation measures: Network cohesion, number of services, service interdependences, absolute importance of a service, absolute dependence of a service, absolute criticality, overall reliability, sizes of input and output messages, message rates, network load, CPU usage, memory usage, average response time, network bandwidth, virtual users

Results: Textual search and 2D semantic search improved with refinements regarding facets and labels included on results. Also, new pagination results solution and specific metadata fields configuration. Profiling tests reveal too many marshal/unmarshal process of data which could affect the query response time.

Conclusions: The inherent complexity of the project integration has been solved successfully with the SCA integration approach, but the downside of it is that performance is not completely satisfactory.

Recommendations: None

5.1.2 General Public and General Practitioner Use Cases

The global computational evaluation for the general public and general practitioner use cases considers the role of the spelling corrector, the translation component and the annotation component on the retrieval process. In so doing we examine the quality of each of these components in isolation and then explore how they impact retrieval. We investigate if there are any parallels in this process, that is parallels between the performance of components in isolation and on how they impact retrieval. The core test set used for these evaluations consists of the Khresmoi document collection and 150 queries with corresponding result set. The query set is comprised of 50 general practitioner queries, 50 short (1–2 word) general public queries and 50 long (>2 words) general public queries. Full details on this evaluation are available in the appended article.

5.1.3 Radiology Use Case

The global computational evaluation for the radiology use case considers the role of the query translation, spelling correction, 2D image search, and relevance feedback (RF) in the search for radiology images using both text and visual queries. The dataset used as a basis for this evaluation is the ImageCLEF 2013 medical image-based retrieval dataset. The dataset contains more than 300.000 images and 35 query topics. It already contained French and German versions of the query set, while the Czech version was created by manual translation of the English topics. Runs were done with queries automatically translated into English (if necessary), with spelling errors artificially introduced and with the introduced spelling errors corrected by the spelling correction service. Full details are available in the appendix in Section 9.

5.2 Interactive System-Level Evaluations

For the interactive system-level evaluation, the three faces of the Khresmoi system were evaluated by end users belonging to the target groups: members of the general public, physicians and radiologists. These systems were evaluated in WP10. Sections 5.2.1 and 5.2.2 provide a meta-analysis of D10.1, which presents the results of the general public and physician interactive evaluations of *Khresmoi for Everyone* and *Khresmoi Professional*. Section 5.2.3 provides a meta-analysis of the interactive evaluations of the *Khresmoi Radiology* system reported in D10.2.

5.2.1 Evaluation of Khresmoi for Everyone

Evaluation points: E1.0, E1.1, E1.2, E1.6, E1.7, E1.8, E1.10, E1.12

Khresmoi tasks: T10.1

Relevant deliverables and sections: D10.3

Resources used/created: Khresmoi For Everyone index (annotated and classified crawled web pages, other metadata used for indexing), user tasks, questionnaires, supporting materials.

Description of experiment(s): The general public user tests reported in this deliverable consist of two parts: A blind and a non-blind comparison of search results between Khresmoi for Everyone and Google, in which the preferences of a total of 22 users completing the online survey were evaluated during December 2013 and February 2014, as well as the results of the large scale user tests conducted in the period from May 2014 to June 2014 conducted on a larger population (N=63 for the general public user tests). The participants were asked to perform search tasks that were described to them, mostly looking for information and using certain capabilities of the system. A questionnaire was given to them afterwards to assess the system efficiency and usability, users satisfaction, etc.

Evaluation measures: User satisfaction (SUS questionnaire, audio-recording, task-related user feedback), Effectiveness (ability of users to solve tasks using the prototype and find the required information/resource), Efficiency (speed and screen recordings showing the effort to find the information), System usability: interface and usability of tools (SUS questionnaire, task-related feedback, amount of training required, screen recordings)

Results: The evaluation was firstly a success in terms of participants recruitment, as it increased considerably since the previous full user tests. Overall, in comparison with the user tests conducted in 2013 we can see a substantial progress regarding evaluation setup, protocol and prototype. The K4E prototype in year 4 was well advanced in terms of user experience and functionalities. In the tests, the sample of participants was very representative of online health information seekers, i.e. internet users of both genders, all ages, different education levels and professional backgrounds. They also varied in their web search experience and skills and personal health experience. Overall, participants were very positive about the prototype, they would not need the help of a technical person to use it. The blind /

non-blind user tests show that presenting K4E as a search engine offering trustworthy health web sites did affect the choice of participants.

Conclusions: All evaluation tests made provided a significant feedback on the acceptance of K4E. A few improvements and fine tuning of the prototypes are still needed on the interface level before the search engine is launched online. Currently, the priority is to increase the coverage of resources in Czech. Following that, our directives will be focused on the improvement of the automatic detection of the readability and trust, and prototype improvements of the interface.

Recommendations: D7.2 recommendations for the evaluation of Khresmoi for Everyone prototype were the following: "*As stated in conclusions, larger samples of end users need to be taken to be able to apply results and conclusions to the larger population. If full tests are being conducted then a near complete working prototype should be used and each core user interface component (e.g. spelling correction, translation) should be evaluated in a separate test.*". These recommendations have been followed: the user evaluation of the prototype has been conducted with a higher number of participants. Moreover, core individual components have been evaluated individually (as described in Section 2). A full list of recommendations based on user's comments and analysis of the results is given based in D10.3, section 3.6.

5.2.2 Evaluation of Khresmoi Professional

Evaluation points: E1.0, E1.1, E1.2, E1.6, E1.7, E1.8, E1.10, E1.12

Khresmoi tasks: T10.1

Relevant deliverables and sections: D10.3

Resources used/created: Khresmoi Professional index (annotated and classified crawled web pages, other metadata used for indexing), user tasks, questionnaires, supporting materials.

Description of experiment(s): Medical professionals were approached at medical education conferences in order to test the Khresmoi Professional search system (Java, web browser or mobile version). The user tests were conducted at four different locations in Year 4: Graz, Wiesbaden, Vienna and online. The number of test sessions included in the evaluation (excludes approx. 10% which had to be removed) is the following: 33 in Graz (November 2013), 16 in Wiesbaden (May 2014), 6 in Vienna (June 2014), 29 online (June/July 2014). A questionnaire was designed to evaluate the usage of tools, search preferences, efficiency and effectiveness of the search system and made accessible in German and English. Participants completed the questionnaire after doing a "free browsing task" or a "pre-defined task" (asking the participant to find information related to a specific query).

Evaluation measures: User satisfaction (SUS questionnaire, audio-recording, task-related user feedback), Effectiveness (ability of users to solve tasks using the prototype and find the required information/resource), Efficiency (speed and screen recordings showing the effort to find the information), System usability: interface and usability of tools (SUS questionnaire, task-related feedback, amount of training required, screen recordings)

Results: The second round of evaluation in Y4 shows improvement over the first round (spring-summer 2014 vs fall 2013), and in comparison with evaluations reported in D10.1 (mid-project).

Conclusions: The report in Section 2 of D10.3 concludes on substantial improvements of the system: "*Unbiased, EU-supported, multilingual access and the availability of Khresmoi functionalities such as the personal library, search facets and summary translation received excellent feedback.*" Requirements from users are related to improvement of the ranking, extended access to pdf and clinically relevant resources, simplification of usability features, advanced integration of tools, implementation of popular functionalities in the web browser, and expansion of facets' content. The analysis of different age and occupational subgroups showed different needs of the users. Self-employed practitioners, in particular, would profit from quicker accessibility to clinically relevant

resources, and older, less IT-adept physicians would appreciate a simplification of usability features. The main finding was related to accessibility issues, as most physicians access medical information via mobile, in particular iOS devices. This means that the Khresmoi Professional search system must be made available on mobile and iOS devices.

Recommendations: It was suggested in D7.1.2 to "*use component-based evaluation, if possible, to assess impact of parts of the system on the overall user satisfaction*". Due to the strategy chosen of creating shorter test sessions to allow them to be carried out at events attended by physicians in order to increase the number of end users participating in the evaluation, this recommendation was not followed. It remains an interesting area to investigate.

5.2.3 Evaluation of Khresmoi Radiology

Evaluation points: E2.0, E2.2, E2.3, E2.5, E2.7, E2.9, E2.11a, E2.11b

Khresmoi tasks: T10.4

Relevant deliverables and sections: D10.4

Resources used/created: ImageCLEFmed datasets and internal data from the MUW PACS, user tasks, questionnaires, supporting materials.

Description of experiment(s): The experiments followed the following plan:

1. Introduction to Khresmoi, the search system and the user test goals (5 minutes)
2. Demonstration of the system functionalities (5 minutes)
3. Demographic survey (5 minutes)
4. Introductory (tutorial) task (5 minutes)
5. User tests using predefined tasks, testing both the 2D and 3D aspects if the system (30-40 minutes)
6. Survey on the satisfaction of the tools and functionalities (10 minutes)
7. Free possibility to use the system (5+ minutes)
8. Survey on the satisfaction with the system, free discussion (10 minutes)

The system was tested with 26 radiologists in four hospitals (Medical University of Vienna, Austria; University Hospitals of Geneva, Switzerland; University Hospital of Freiburg, Germany; and University Hospital of Larissa, Greece). The average duration of the tests was 65 minutes. Comparisons were made with the results of the user tests on the year 2 prototype and the year 1 survey self assessments on the current image search.

Evaluation measures: Success in information finding, speed in information finding, learnability and user satisfaction (using modified versions of the SUS and QUIS)

Results: A high success rate in information finding using the system was achieved (100% for the 2D image-based tasks compared to 81% in the year 2 evaluation; 85% for the article search task, compared to 79% in the year 2 evaluation; 86% for the 3D tasks) in a low time frame (average of 102 and 86 seconds for the two 2D image-based tasks, compared to 106 seconds in the year 2 evaluation; similar times for the article search task in the year 2 and year 4 evaluations (150 and 159 seconds respectively); average of 260 seconds for the 3D tasks in the year 4 evaluation). The response times were now acceptable for the 3D prototype in the year 4 evaluations, with a reduction from over a minute between pressing the search button and being able to interact with results in year 2, to a 3 second query response time and 17 seconds to load the volume details in year 4. The user satisfaction scores were higher than the first evaluation round, especially in low performing aspects such as response speed (Median Likert value increased from 3 in the year 2 user evaluation to 5 in the year 4 evaluation for the 3D prototype), result satisfaction (Median Likert value increased from 2 in the year 2 user evaluation to 4 in the year 4 evaluation for the 3D prototype) and consistency of the complete

integrated system (Median Likert value increased from 3 in the year 2 user evaluation to 4 in the year 4 evaluation). Comments on improvements and wishes for additional features were collected.

Conclusions: The overall user satisfaction was positive. Radiologists expressed belief that they would use the system in academic, research and clinical duty. The newly added tools achieved high scores in terms of user satisfaction. Aspects of the prototype interface that need improvement or modification in a potential commercial product were identified. 3D CBIR retrieval may have difficulties when multiple pathologies exist in a marked ROI. Modality classification in 2D retrieval, though useful, needs to be more accurate. Semantic image search can be useful if interfaced in a less explicit way (i.e. requiring less knowledge about semantic queries from the end users).

Recommendations: It would be useful to compare the Khresmoi system with other systems such as Goldminer and Yottalook in the same user-centred evaluation framework. For the evaluation of the 3D framework, it would be useful to measure how many relevant results were found.

5.3 Summary and Analysis

The performance analysis of the Khresmoi architecture in Section 4.1 shows that the choice of an SCA approach allowed effective integration of multiple components in the Khresmoi Cloud, although the performance provided by this approach is not completely satisfactory.

Two computational system-level evaluations investigated the impact of separate components on the full systems, described in Section 4.2. For the general public and physician use cases, the impact of the spelling corrector, the translation component and the annotation component on the retrieval process was investigated. For the 2D radiology image retrieval, the impact of the use of the spelling corrector as well as various relevance feedback approaches was investigated. For the 2D radiology image retrieval, the result that introducing spelling errors has different impact in different languages was obtained, as was the result that automatically correcting the spelling has a positive impact on all four languages, but a different magnitude of impact for each language. The most successful relevance feedback approach was using the fusion of text and visual information and late fusion for fusing multiple image queries as well as text with visual results (the mixed_lf approach).

Section 4.3 provides details on the full user tests which were conducted for the general public, medical practitioners and radiologists. The methodology adopted for these final user tests has been streamlined and is also more harmonized over the three use cases than for the previous user-centred evaluations. The emphasis on developing rigorous protocols taking into account the experience in the year 2–3 evaluations led to well conducted experiments.

The main shortcoming related to the interactive evaluations identified in the previous meta-analysis (D7.2) was the small number of participants in the evaluations (small population size). Attracting a sufficiently large and representative population of users was made more challenging by the insistence in Khresmoi on using “real” physicians and radiologists, and not surrogates such as medical students. For the general public, the year 4 evaluations had 63 participants, compared to 28 participants for the year 2 evaluations. 84 physicians participated in the year 4 evaluations, compared to 14 participants in the year 2 evaluations. Finally, 26 radiologists participated in the year 4 evaluations, compared to 17 radiologists in the year 2 evaluations. For the general public and radiology use cases, the additional participants were mainly obtained by increasing effort in recruitment and conducting the evaluations in multiple countries and locations. The large increase in participation in the physician use case is due to the new approach developed and adopted after the difficulties faced in recruiting participants in the year 2 evaluations. Conducting the evaluations at physician symposia proved to be an effective way to attract participation from the target group, however the increased participation had to be balanced by a maximum evaluation session length of 20 minutes at such an event. In order to obtain the maximum amount of information in such a short time, the experimental protocols were optimised e.g. by avoiding asking effectively the same question multiple times phrased in different ways.

The interactive evaluations were facilitated in year 4 by the maturity of the prototypes after three-and-a-half years of development. This allowed the use of open tasks in the evaluation protocols that allowed useful information to be gathered on the normal search behaviour of the end users.

For each of the three prototypes, user feedback was generally positive, although areas for improvement were also identified. Advances made in the methodology for conducting interactive evaluations, in particular for the physician use case, will be useful for future evaluations. Nevertheless, further improvements and streamlining are potentially possible to allow even more efficient experiments.

6 Conclusion

This deliverable presents a meta-analysis of four types of evaluation conducted in the second and final round of evaluations in the Khresmoi project: computational component-level, computational system-level, interactive component-level and interactive system-level evaluations. Overall we find that high quality evaluations and research have been conducted in the project.

In the Khresmoi project, particular emphasis was placed on a coordinated and independently validated evaluation strategy. The initial version of the evaluation strategy was made available to the consortium in D7.1.1, and a meta-analysis of the results obtained in the evaluations based on this strategy was provided in D7.2. This exercise proved to be very useful, as four major shortcomings and a number of shortcomings specific to individual experiments were identified. The updated evaluation strategy made available to the consortium as D7.1.2 suggested approaches to overcome these shortcomings. The final document in the series is D7.3 (this document), which presents the meta-analysis of the second and final round of evaluations conducted in the Khresmoi project, based on the updated strategy.

A change in the methodology was adopted for the creation of this document compared to the previous meta-analysis (D7.2). For the previous meta-analysis, the authors of the meta-analysis took on the task of searching through the evaluation deliverables published by the various project work packages for the information needed for the meta-analysis. For this second round of the meta-analysis, an information gathering form (Figure) was distributed before the evaluations were conducted to the groups conducting them. The form served two main purposes: (i) it reminded the groups about the important aspects of the evaluation, and (ii) it allowed the meta-analysis to be created more efficiently as the pertinent information and references to the relevant sections of the deliverables were captured in the forms.

The four main shortcomings identified in the previous meta-analysis (D7.2) have all been overcome to a large extent, as described below:

Data: New data sets have become available through evaluation campaigns either organised by Khresmoi members or externally. There are still some deficiencies in the evaluation data available for the text analysis and retrieval evaluations that need to be resolved beyond Khresmoi.

Evaluation protocol: More attention was paid to the metrics to be calculated, although there was still a tendency to omit providing results for the efficiency metrics. There was also a better comparison of the Khresmoi approaches to the state-of-the-art.

Component performance: No components were identified to be performing insufficiently. However, it was found that the overall efficiency of the integrated system based on the SCA architecture was not as good as expected.

Population size: All of the interactive evaluations had sufficient participants in the second round of evaluations. An impressive increase in the number of medical professionals participating in the interactive evaluation of Khresmoi Professional was attained by adopting the approach of doing evaluations at physician congresses. This increase in the number of participants was balanced by a

necessary reduction in the time spent by each participant on evaluating the system, but effort was made to streamline the evaluation protocol so as to obtain as much data as possible in the short time available.

The adopted approach of examining the performance at both component-level and system-level, as well as with interactive and computational experiments, gave useful insights into optimising the overall system design. It was noted, for example, that even though the spell checkers often seem to damage the queries themselves (Section 4.1.3.3), their use to correct misspelled queries submitted to a search engine does lead to more relevant results (Section 5.1.3). Similarly, input from physicians participating in interactive evaluations led to the identification of additional resources to index and hence improved the ability of the system to provide relevant results to certain queries.

The approach adopted to implement this coordinated evaluation strategy was heavily manual, but given that this is the first time that such a comprehensive evaluation strategy was implemented for a domain-specific search system, this could not be avoided. The participants gained valuable knowledge and experience in doing this, which was particularly noticeable in improvements in the procedure between the first and second rounds of meta-analysis. The fact that the first meta-analysis identified key shortcomings that were overcome during the second round of evaluations, in particular in a very creative way for the interactive evaluation of Khresmoi Professional, supports its usefulness.

A promising way forward is to automate more of the evaluation process. Given an architecture for component integration such as the one adopted in Khresmoi, it would be interesting to run regular automated evaluations of the components and of the combinations of components, similar to the component-level approach adopted in [16]. This could be combined with an approach in which end users have access to the prototypes and can carry out searches on them, an approach currently promoted by the Living Labs initiative [17]. Work toward the end of the Khresmoi project on attracting medical professionals to participate in online evaluations and creating networks of these professionals through social media demonstrated the feasibility of Living Labs, even for busy professionals such as physicians.

7 References

[1] D7.1.1 - User-centered and empirical evaluation strategy and recommendations. Confidential Technical Report, Khresmoi Project, 31 August 2011.

[2] VA Zavaletta, BJ Bartholmai, RA Robb, High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. Academic radiology 14(7) (2007) 772–78)

[3] A Hanbury and M Lupu, Toward a Model of Domain-Specific Search, Proc. Open Research Areas in Information Retrieval (OAIR) Conference, 2013, pages 33–36, Lisbon, Portugal.

[4] M Salampasis and A Hanbury, A Generalized framework for Integrated Professional Search Systems, Proc. IRF Conference, 2013, Springer LNCS 8201, pages 99–110.

[5] A Hanbury and H Müller, Automated Component-Level Evaluation: Present and Future, Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation 2010, Springer LNCS 6360, pages 124–135, Padua, Italy.

[6] L Goeuriot, L Kelly, GJF Jones, G Zuccon, H Suominen, A Hanbury, H Müller and J Leveling, Creation of a New Evaluation Benchmark for Information Retrieval Targeting Patient Information Needs, Proc. 5th International Workshop on Evaluating Information Access (EVIA), 2013, Tokyo, Japan.

[7] A Hanbury, H Müller, G Langs, M Weber, BH Menze and T Salas Fernandez, Bringing the Algorithms to the Data: Cloud-based Benchmarking for Medical Image Analysis, In Proc. of the CLEF Conference, 2012, Springer LNCS 7488, pages 24–29, Rome, Italy.

[8] R Donner et al. Evaluation of fast 2D and 3D medical image retrieval approaches based on image miniatures. Proc of MICCAI'11, Medical Content-Based Retrieval for Clinical Decision Support, pp.128-138

[9] R Donner et al. Fast Anatomical Structure Localization Using Top-down Image Patch Regression. In Proc. of MICCAI 2012 Workshop on Medical Computer Vision

[10] R Donner, et al. Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Medical image analysis, 17(8):1304–1314, 2013.

[11] D Kelly, Methods for Evaluating Interactive Information Retrieval Systems with Users, Foundations and Trends in Information Retrieval, Vol. 3, Nos. 1–2, pp. 1–224, 2009.

[12] L Goeuriot, GJF Jones, L Kelly, J Leveling, A Hanbury, H Müller, S Salanterä, H Suominen, G Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In Proceedings of CLEF 2013, online working notes.

[13] O Bojar, C Buck, C Federmann, B Haddow, P Koehn, J Leveling, C Monz, P Pecina, M Post, H Saint-Amand, R Soricut, L Specia, A Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, p. 12-58.

[14] L Goeuriot, O Hamon, A Hanbury, GJF Jones, L Kelly, J Leveling, J Robertson, A Tamchyna, D7.2 – Meta-analysis of the first phase of empirical and user-centered evaluations, Technical Report, Khresmoi Project, August 2013.

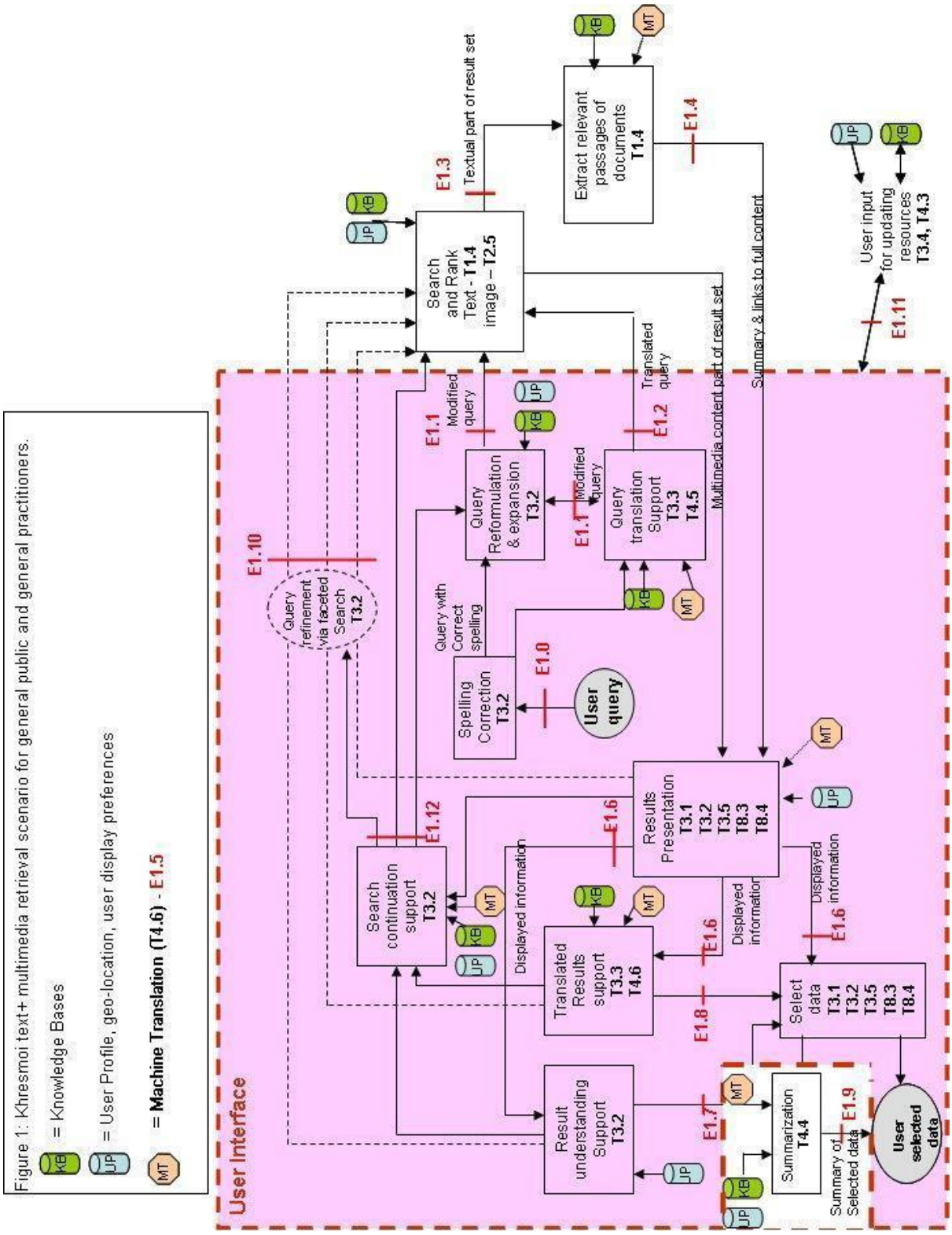
[15] D7.1.2 – Updated user-centered and empirical evaluation strategy and recommendations. Confidential Technical Report, Khresmoi Project, 20 December 2013.

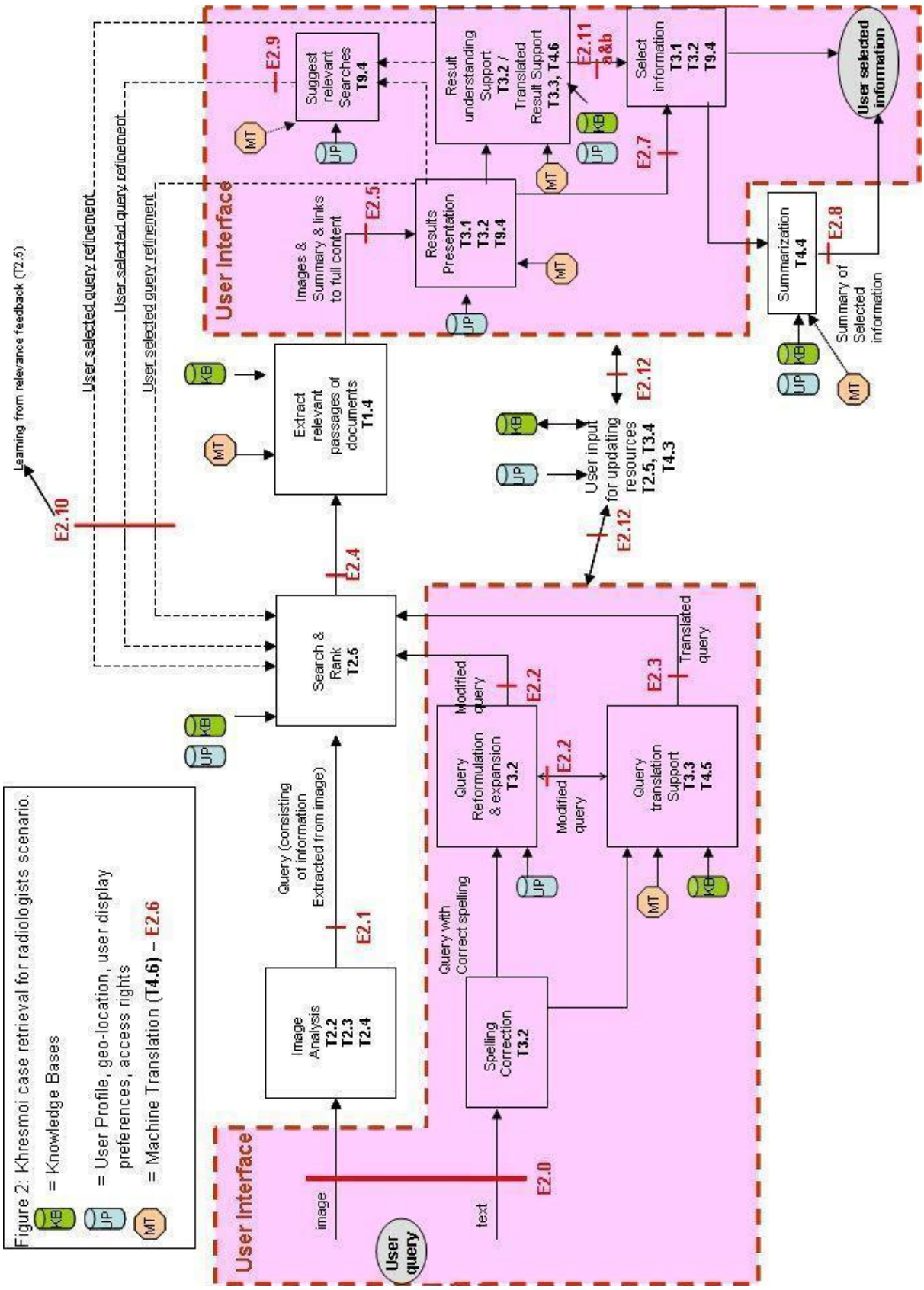
[16] J Kürsten, M Eibl, A Large-Scale System Evaluation on Component-Level, Proc. ECIR 2011, Springer LNCS 6611, pp. 679–682, 2011.

[17] K Balog, L Kelly, A Schuth, Head First: Living Labs for Ad-hoc Search Evaluation, 23rd ACM International Conference on Information and Knowledge Management (CIKM '14), November 2014, to appear.

8 Appendix: Evaluation Diagrams

The appendix contains a copy of the diagrams on pages 8–10 of D7.1.2 for ease of reference.




User Interface
User query
image
text
E2.0
Image Analysis T2.2 T2.3 T2.4
Query (consisting of information Extracted from image)
E2.1
Spelling Correction T3.2
Query with Correct spelling
Query Reformulation & expansion T3.2
Modified query
E2.2
Query translation Support T3.3 T4.5
Translated query
E2.3
Search & Rank T2.5
E2.4
Extract relevant passages of documents T1.4
Images & Summary & links to full content
E2.5
Results Presentation T3.1 T3.2 T9.4
Result understanding Support T3.2 / Translated Result Support T3.3, T4.6
E2.11 a&b
Suggest relevant Searches T9.4
E2.9
Select information T3.1 T3.2 T9.4
E2.7
User selected information
Summarization T4.4
E2.8
Summary of Selected information
User input for updating resources T2.5, T3.4 T4.3
E2.12
E2.12
E2.10
Learning from relevance feedback (T2.5)
User selected query refinement
User selected query refinement
User selected query refinement
KB
UP
MT
Figure 2: Khresmoi case retrieval for radiologists scenario.
= Knowledge Bases
= User Profile, geo-location, user display preferences, access rights
= Machine Translation (T4.6) – E2.6

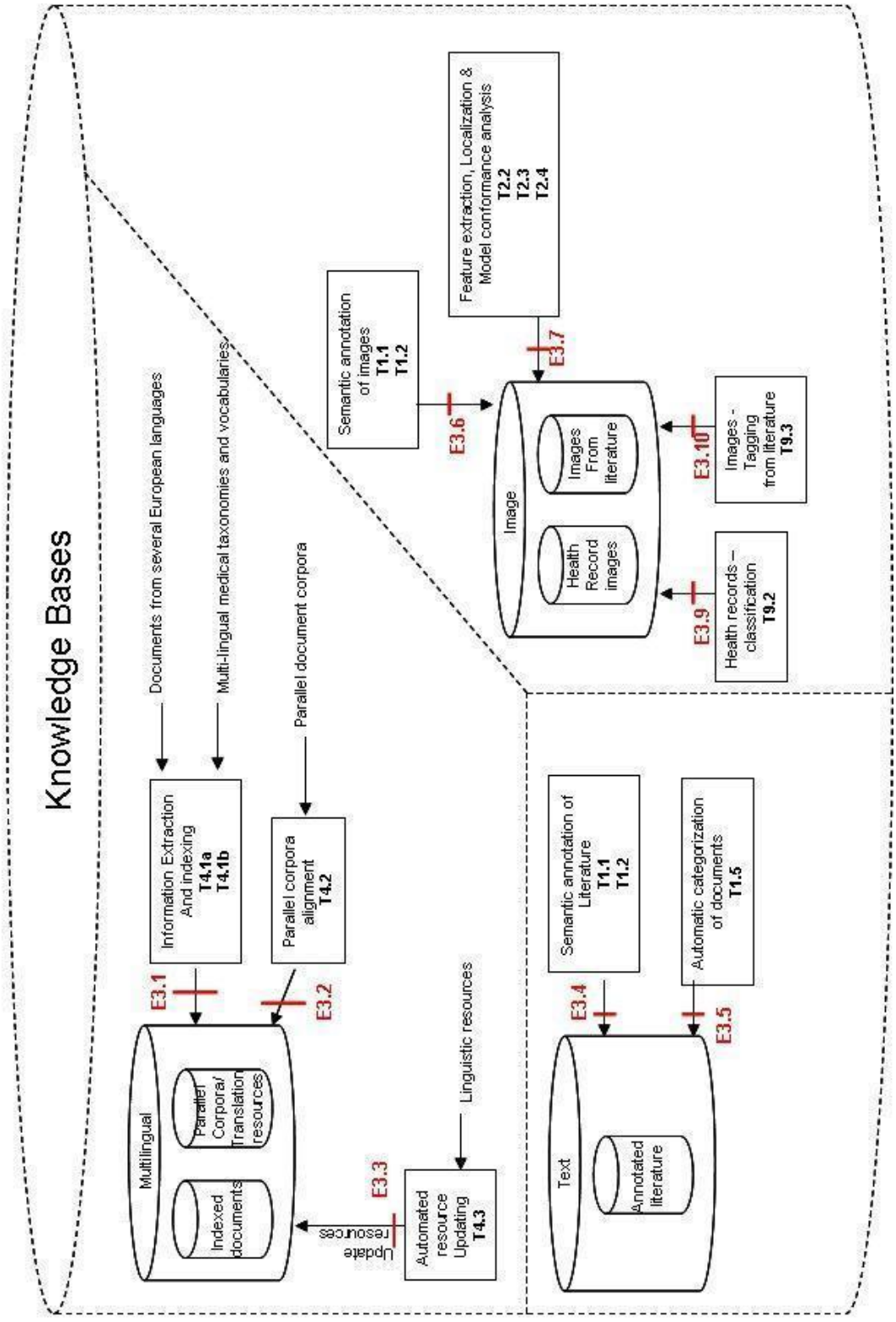


Figure 3: Khresmoi Knowledge Bases.

9 Appendix: Radiology Use Case Computational System-Level Evaluation

In this section the global evaluation of the Radiology prototype is presented. The goal is to assess the role the various Khresmoi components play in the search pipeline, from the query entered by the user until the end of the search. For this purpose we evaluate the effect of the following Khresmoi components: query translation, spelling correction, 2D image search and relevance feedback (RF).

9.1 Test Set

The dataset used as a basis for this evaluation is the ImageCLEF 2013 medical image-based retrieval dataset. The dataset contains more than 300.000 images and 35 query topics. It already contained French and German versions of the query set, while the Czech version was created by manual translation of the English topics.

9.2 Evaluation Approach

The same strategy was followed as with the global empirical evaluation described in Khresmoi deliverable D7.2, Section 4.1. Three main runs were evaluated for each language:

- The first one used the queries after being translated into English by the query translation service, described in Khresmoi deliverable D4.5.
- The second one used the queries after being translated into English by the query translation service and spelling errors were artificially introduced as described in D7.2.
- The third one used the translated queries as the two runs above with the spelling errors corrected by the spelling correction service, as described in D7.2.

All the above queries were used as input for the 2D image search engine. A similar experiment to the one used for the relevance feedback evaluation described in Khresmoi deliverable D2.7 was set up. All the Relevance Feedback (RF) techniques supported by the Khresmoi 2D image search engine were evaluated. Four search iterations were simulated and the top 100 results were used for the RF. The five RF techniques supported are briefly described in the following list (more details can be found in D2.7):

- text: RF using only text information.
- visual_lf: RF using only visual information and late fusion for fusing multiple image queries.
- visual_rocchio: RF using only visual information and the Rocchio algorithm for fusing multiple image queries.
- mixed_lf: RF using fusion of text and visual information and late fusion for fusing multiple image queries as well as text with visual results.
- mixed_rocchio: RF using fusion of text and visual information, the Rocchio algorithm for fusing multiple image queries and late fusion for fusing of text with visual results.

9.3 Results and Analysis

The mean average precision (MAP) values of the runs are presented in Tables 1 - 12. In Tables 1-4 no spelling errors are introduced, thus evaluating the effect of the translation component in the retrieval performance.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.17834	0.17834	0.17834	0.17834	0.17834
1	0.33700	0.23378	0.34676	0.21345	0.33622
2	0.3694	0.23262	0.37782	0.21316	0.37195
3	0.37392	0.23262	0.38104	0.21316	0.3752
4	0.37417	0.23262	0.38110	0.21316	0.37535

Table 2. MAP values per search iteration for the supported RF techniques. They are English runs that contain no introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.12697	0.12697	0.12697	0.12697	0.12697
1	0.21802	0.1691	0.22152	0.16036	0.2134
2	0.2534	0.17566	0.25841	0.16325	0.25723
3	0.26364	0.17576	0.27296	0.16325	0.27218
4	0.2678	0.17576	0.27479	0.16325	0.27378

Table 3. MAP values per search iteration for the supported RF techniques. They are German runs whose queries were translated into English using the query translation service. They contain no introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.16833	0.16833	0.16833	0.16833	0.16833
1	0.29388	0.21001	0.30768	0.19368	0.29426
2	0.32997	0.21021	0.34603	0.19469	0.33964
3	0.33861	0.21018	0.35475	0.19359	0.34866
4	0.34282	0.21018	0.35682	0.19359	0.35084

Table 4. MAP values per search iteration for the supported RF techniques. They are French runs whose queries were translated into English using the query translation service. They contain no introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.13817	0.13817	0.13817	0.13817	0.13817
1	0.27701	0.17971	0.28873	0.16202	0.28001
2	0.31358	0.17862	0.3185	0.16178	0.31396
3	0.31525	0.17862	0.32127	0.16178	0.31663
4	0.31555	0.17862	0.32153	0.16178	0.31693

Table 5. MAP values per search iteration for the supported RF techniques. They are Czech runs whose queries were translated into English using the query translation service. They contain no introduced spelling errors.

Judging by the MAP values in the iteration 0 in the Tables 1-4 it can be derived that the error introduced by the French to English translation is the smallest (MAP score of 16.8% compared to 17.8%), followed by the translation Czech to English (MAP score of 13.8%). The German to English translation introduced the biggest drop in retrieval performance (MAP score of 12.7%).

The MAP values in search iteration 4 show the RF added value in refining a search, improving the performance in all of the cases, with the mixed RF techniques performing slightly better than the text only RF techniques and visual only RF techniques achieving the worse performance. These results confirm the findings of D2.7.

Tables 5-8 contain MAP scores for the translated queries with spelling errors introduced artificially. Iteration 0 shows similar performance of translated German and Czech runs to the English run (MAP score of 4.5% and 3.2% compared to 4.7%) when they contain spelling errors, while the French run performs even better than the English one (MAP score of 9%).

Then Tables 9-12 present the retrieval performance results after the error-containing queries are automatically corrected by the spelling correction service. The spelling correction component always improves performance over the spelling erroneous runs and search refinement by RF by all methods improves retrieval scores in all cases as well. The mixed and text-only runs achieve the best performance again.

Figure 1 shows a summary diagram for the mixed_lf iteration 4 values taken from Tables 1–12. Here it is clear that the French runs are affected to a lesser extent than the other languages by the introduced spelling errors, but also that the spelling correction has the lowest improvement for French, while resulting in noticeable improvements for English, German and Czech.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.04679	0.04679	0.04679	0.04679	0.04679
1	0.12357	0.08867	0.12183	0.08105	0.11895
2	0.15219	0.0883	0.15457	0.07899	0.14705
3	0.16114	0.08855	0.16762	0.07882	0.15820
4	0.16313	0.08855	0.17034	0.07882	0.16287

Table 6. MAP values per search iteration for the supported RF techniques. They are English runs that contain artificially introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.04481	0.04481	0.04481	0.04481	0.04481
1	0.10525	0.07445	0.10683	0.06525	0.10405
2	0.10223	0.07552	0.11711	0.06149	0.11390
3	0.10758	0.07552	0.1246	0.06149	0.1233
4	0.11091	0.07552	0.12768	0.06149	0.12388

Table 7. MAP values per search iteration for the supported RF techniques. They are German runs whose queries were translated into English using the query translation service. They contain artificially introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.09040	0.09004	0.09040	0.09040	0.09040
1	0.19279	0.14242	0.19345	0.12712	0.17923
2	0.20896	0.13984	0.21569	0.12355	0.19587
3	0.21588	0.13995	0.22178	0.12359	0.20161
4	0.21737	0.14001	0.22498	0.12363	0.20498

Table 8. MAP values per search iteration for the supported RF techniques. They are French runs whose queries were translated into English using the query translation service. They contain artificially introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.03213	0.03213	0.03213	0.03213	0.03213
1	0.10922	0.05294	0.11678	0.04139	0.10844
2	0.11704	0.05286	0.12661	0.04135	0.11953
3	0.12402	0.05287	0.13358	0.04136	0.12727
4	0.12614	0.05287	0.13550	0.04136	0.12892

Table 9. MAP values per search iteration for the supported RF techniques. They are Czech runs whose queries were translated into English using the query translation service. They contain artificially introduced spelling errors.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.13493	0.13493	0.13493	0.13493	0.13493
1	0.22543	0.18428	0.22644	0.17177	0.21487
2	0.24548	0.18573	0.25412	0.17004	0.23896
3	0.24862	0.18339	0.25683	0.17004	0.24169
4	0.24881	0.18339	0.25688	0.17004	0.24178

Table 10. MAP values per search iteration for the supported RF techniques. They are English runs that contain artificially introduced spelling errors, corrected by the HON spell correction service.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.10234	0.10234	0.10234	0.10234	0.10234
1	0.21347	0.14661	0.21779	0.13330	0.21348
2	0.24174	0.1449	0.24714	0.13095	0.24668
3	0.24767	0.14482	0.25561	0.13094	0.25437
4	0.25018	0.14482	0.25615	0.13094	0.25475

Table 11. MAP values per search iteration for the supported RF techniques. They are German runs whose queries were translated into English using the query translation service. The spelling errors were corrected by the spelling correction service.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.09733	0.09733	0.09733	0.09733	0.09733
1	0.19354	0.13713	0.20473	0.12238	0.19881
2	0.20822	0.13748	0.22015	0.12322	0.21310
3	0.21287	0.13748	0.22464	0.12215	0.21674
4	0.21339	0.13748	0.22579	0.12215	0.21812

Table 12. MAP values per search iteration for the supported RF techniques. They are French runs whose queries were translated into English using the query translation service. The spelling errors were corrected by the spelling correction service.

Method/iteration	text	visual_lf	mixed_lf	visual_rocchio	mixed_rocchio
0	0.09951	0.09951	0.09951	0.09951	0.09951
1	0.21590	0.16174	0.22582	0.15545	0.20333
2	0.24259	0.16328	0.25312	0.14919	0.23708
3	0.24916	0.16328	0.26117	0.14919	0.24480
4	0.25181	0.16328	0.26321	0.14919	0.24664

Table 13. MAP values per search iteration for the supported RF techniques. They are Czech runs whose queries were translated in English using the query translation service. The spelling errors were corrected by the spelling correction service.

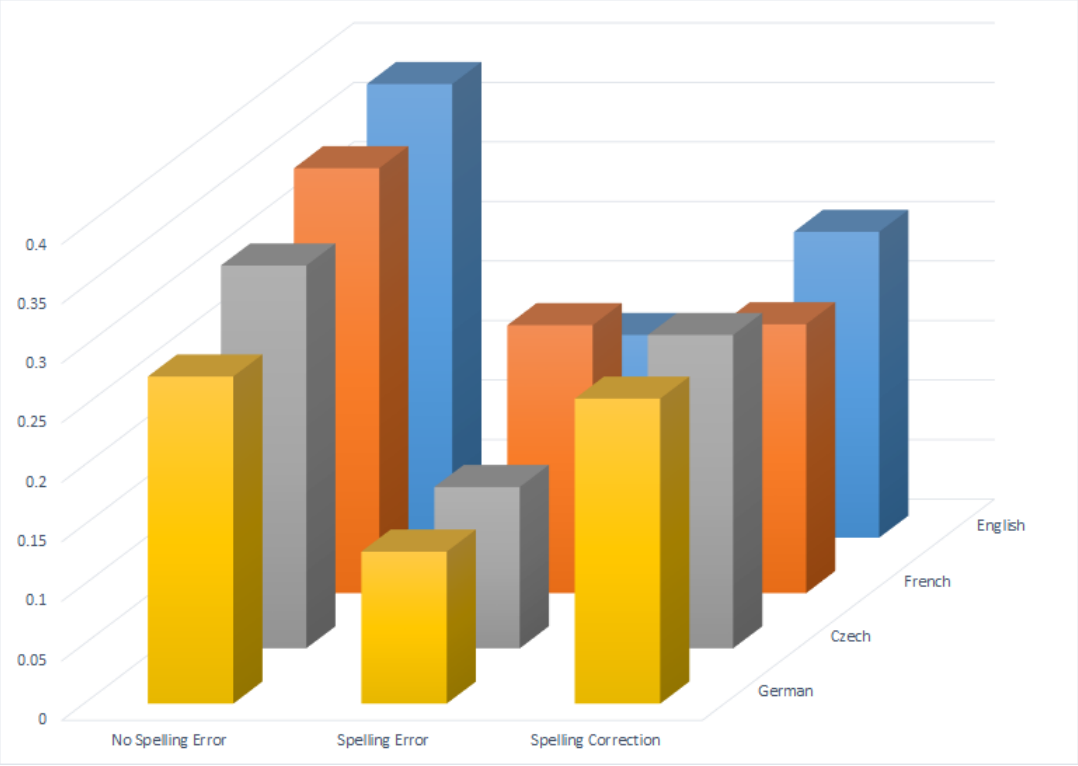


Figure 1. Plot of the mixed_lf values for iteration 4 for each of the languages (i.e. one value from each of Tables 1-12). The values are for, respectively, no spelling error (Tables 1-4), spelling error introduced (Tables 5-8) and spelling correction applied (Tables 9-12). Machine translation into English is applied for the French, Czech and German queries.