Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on results of the WP1 second evaluation phase

Deliverable number	*D1.8*
Dissemination level	*Public*
Delivery date	*June 2014*
Status	*Final*
Author(s)	*Wei Li, Angus Roberts, Johann Petrak, Ljiljana Dolamic, Gareth J.F. Jones, Liadh Kelly, Lorraine Goeuriot*

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

Executive Summary

This report describes evaluations conducted on semantic annotation, categorisation, and recommender systems in the Khresmoi project.

By semantic annotation, we mean the linking of spans of texts to semantic information held in a knowledge base. Indexing of such semantic annotations allows for richer search, and such indexes underlie several of the Khresmoi search interfaces. This report examines the intrinsic quality of the Khresmoi semantic annotations by (a) summarising evaluations reported elsewhere in Khresmoi, and (b) providing a new comparison, to a resource external to Khremsoi, the CALBC Silver Standard.

The report also discusses extension of the Khresmoi web page categorization system for automatic detection of HONCode principle for other languages besides English, based on the work in deliverable 1.6. The categorization system was also integrated into the crawler pipeline for these languages. Czech is also implemented into the K4E semantic search interface in this deliverable.

Finally, we investigate the effectiveness of an integrated model on a small part of the data collection in this deliverable. This integrated model combines the standard IR system with a recommender component to generate a combined output for users based on previous users' search behaviour. In this deliverable, the content-based filtering method is chosen for the recommender component to compute the prediction for users. And obtained results show that this integrated model has the potential to improve the retrieval results to better satisfy users' information needs.

All components perform their role satisfactoraly. Section 2 and 3 contain an extension of methodological described in previous deliverables. And section 4 presents a novel retrieval method for the project.

Table of Contents

List of Figures

List of Tables

List of Illustration

1 Introduction

This report presents results on the evaluation of technologies created in Khresmoi Workpackage 1 for the enhancement of user search. These include semantic annotation, web page categorisation, and search results recommendation.

Section 2 examines the intrinsic quality of the semantic annotation used in several of the Khresmoi search prototypes. By intrinsic quality, we mean the quality of the semantic annotations in themselves, rather than the usefulness of such annotation, which is reported in the Khresmoi user evaluations. We provide background to the semantic annotation task, summarise a manual evaluation of the Khresmoi annotations, and provide a comparison to a third-party, non-Khresmoi annotation resource, the CALBC Silver Standard. This latter evaluation is important in that it places the Khresmoi semantic annotation system in the context of other annotation systems, and provides useful insights in to the correctness of the Khresmoi system.

The categorization of the web pages according to 8 HONcode principles performed by HON was described in deliverable 1.6. In deliverable 1.6 this categorization was evaluated for English. In this deliverable, besides English, HON evaluates the system for automatic detection of HONcode principles for other languages namely French, German, Spanish, Italian and Dutch. Additionally, this system was integrated into the crawler pipeline for these non-English languages. Implementation of the categorization results into the K4E search system is also described in this document. Implementation of Czech in K4E semantic search interface in addition to English is also presented in this deliverable.

In section 4, a novel retrieval method is proposed and investigated. In order to aid users find better retrieval results, DCU exploit the recorded previous users search logs to generate the prediction results using a collaborative filtering algorithm, and combine this recommendation list with the standard information retrieval (IR) result. An initial experiment is conducted on the integrated retrieval model on a small test collection. This obtained relatively good output and shows it has the potential to further improve the effectiveness of the IR results.

2 Semantic Annotation against Known External Resource

2.1 Chapter introduction

This Chapter describes the intrinsic evaluation of the Khresmoi semantic annotation, i.e. measures of the quality of the annotations. This intrinsic evaluation is presented by reference to other deliverables specific to an evaluation of the semantic annotation, and by detailing a quantitative comparison of the Khresmoi annotation pipeline with trusted third party resources external to the project. Extrinsic evaluations of the semantic annotation, i.e. measures of the usefulness of the annotations in an end-user setting, are not presented here, as these are part of the overall Khresmoi user evaluations.

2.1.1 Background

This section provides a brief summary of the use of semantic annotations in Khresmoi, in order to motivate the analysis given in later sections.

By *semantic annotation* we mean the identification and recording of spans of text as entities that belong to a specific class, and the linking of those entities to some semantic knowledge resource, such as an ontology. In the case of Khresmoi, the entity classes are those identified in Khresmoi Deliverable D1.1 [9], and summarised in Table 1. Linkage is provided to the Khresmoi Knowledge Base [11], by way of Khresmoi class identifiers and instance identifiers. For most of the Khresmoi entities, and specifically for those analysed in the bulk of this report, the linkage is to concepts from UMLS [15], as represented in the Khresmoi Knowledge Base. The class and instance identifiers used by the knowledge base in these cases are the UMLS concept identifier (CUI) and the UMLS type identifier (TUI). TUIs are given for each of the Khresmoi classes in Table 1.

Semantic annotation is carried out by the Khresmoi annotation pipeline, described in Khresmoi Deliverable D1.2 [10]. The pipeline consists of multiple of language processing steps, run in a specific sequence over a document. Each processing step builds on the information generated by previous steps. Steps include:

- basic lexical and syntactic processing, such as tokenisation and part-of-speech recognition;
- handling of document formats;
- metadata extraction;
- lookup of terms in large dictionaries compiled from the Khresmoi knowledge base;
- hand-crafted disambiguation rules;
- machine learned disambiguation models.

Once a document has been annotated, the full text, annotations, and the knowledge base class and instance information are indexed in the Khresmoi index server. They are thus made available for search in the various Khresmoi clients. Clients currently use the semantic information, i.e. the linkage from the indexed entities to the knowledge base, in one of three ways:

1. In the Khresmoi Professional client, user queries are expanded using synonym information and taxonomic relations in the Khresmoi Knowledge Base. Annotation and storage of class and instance identifiers in the Khresmoi text collection allows this expanded query to be executed across the indexed annotations.
2. In the Khresmoi for Everyone client, users can construct queries from concepts provided by the Khresmoi Knowledge Base, and these are again executed across class and instance

identifiers on the indexed annotations. For example, a user might construct a query to find documents mentioning [Diseases] that can be [Treated] with [Anti-inflammatories], where the terms in square brackets represent concepts in the Khresmoi knowledge base, used as

3. In the Khresmoi Radiology client, class and instance identifiers are used to link annotations in radiology reports to annotations in the literature, again via relationships stored in the Khresmoi Knowledge Base.

Khresmoi annotation type	**TUI**	**TUI description**
Anatomy	T029	Body Location or Region
	T023	Body Part, Organ, or Organ Component
	T030	Body Space or Junction
	T022	Body System
	T024	Tissue
	T190	Anatomical Abnormality
	T020	Acquired Abnormality
	T019	Congenital Abnormality
Disease	T033	Finding
	T037	Injury or Poisoning
	T046	Pathologic Function
	T047	Disease or Syndrome
	T048	Mental or Behavioral Dysfunction
	T049	Cell or Molecular Dysfunction
	T050	Experimental Model of Disease
	T184	Sign or Symptom
	T191	Neoplastic Process
Investigation	T060	Diagnostic Procedure
	T059	Laboratory Procedure
	T034	Laboratory or Test Result
	T062	Research Activity
	T063	Molecular Biology Research Technique
Drug	T200	Drug (Clinical Drug)
	T121	Pharmacologic Substance
	T195	Antibiotic

Table 1: Definition of the Khresmoi semantic annotation types, in terms of UMLS CUIs. Note that in addition to UMLS Drugs, drugs found in the DrugBank resource are also annotated by the pipeline. This report, however, does not consider DrugBank annotations, and where Drug is mentioned, this implies Drugs defined in terms of UMLS TUIs.

2.1.2 Summary of this chapter

The following section of this chapter outlines the evaluation metrics used. The two intrinsic evaluations of the semantic annotation that have been carried out in Khresmoi are then described:

[1] A manual evaluation, in which samples of annotations created by the Khresmoi pipeline were examined by trained human annotators, and corrected. The corrections were then compared to the original annotations, to generate metrics of agreement. This is presented in outline only, in Section 2.3.

[2] An automatic evaluation, using a resource previously annotated with known standard annotations. The evaluation consisted of running the Khresmoi pipeline over this standard resource, and comparing Khresmoi annotations with standard annotations. This is fully described in Section 2.4.

2.2 Evaluation metrics

Evaluation metrics used in this chapter are standard measures of precision, recall and agreement, as presented in Khresmoi Deliverable D1.3 "Report on results of the WP1 first evaluation phase" [14]. We adapt the description from D1.3 below, for ease of reference.

We measure performance of automatic systems against standard collections, human correction of the system's output, and against other systems, using standard information retrieval metrics. In each case, we refer to annotations in the standard, or created by human correction, or created by some other system, as the set of "Key" annotations. We refer to annotations created by the system to be compared with these (i.e. the Khresmoi system) as the set of "Response" annotations.

We define:

- *true positive* as an annotation in the Response set that matches an annotation in the Key set.
- *false positive* as an annotation in the Response set that does not match an annotation in the Key set.
- *false negative* as an annotation in the Key set that is not matched by an annotation in the Key set.

Our metrics are then defined as follows:

Precision,

$$P = \frac{true\ positives}{true\ positives + false\ positives}$$

measures the correctness in terms of what percentage of the annotations created in the Response agree with the Key.

Recall,

$$R = \frac{true\ positives}{true\ positives + false\ negatives}$$

measures the coverage of the Response compared to the Key, or what proportion in the Key is in the Response.

We then can calculate the harmonic mean of P and R, the F statistic, which is defined as:

$$F = \frac{2PR}{P+R}$$

More correctly, this is the balanced F measure or F1, where equal weight is given to P and R.

In defining positive and negative classes, and therefore the metrics calculated from them, we may either define metrics as *lenient*, in that we allowed overlapping annotations to be considered matches, or as *strict*, in that we do not consider overlapping annotations to be matches.

Where we are comparing a Khresmoi system Response against a Key from some other system, as opposed to a Key from a standard, we can define the agreement between the systems – this is known as the Inter Annotator Agreement, IAA. We use the method described in [12], summarised here. IAA is calculated from the number of matches and non-matches between the two systems. For every match from the Key, there will also be a match from the Response. The total number of matches is therefore double the number of matches from any one system. The total number of non-matches is the sum of non-matches from each system.

IAA can then be calculated as:

IAA = matches / (matches + non-matches)

IAA can be shown to be equivalent to un-weighted F-measure [13].

2.3 Human correction of Khresmoi annotation

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. This is an important component of the evaluation of the quality of the Khresmoi semantic annotations, and this section therefore summarise the work, references the deliverables in which it is reported, and summarises the main evaluation conclusions of the correction work.

Automatic and corrected annotations were created over the Khresmoi reference corpus, as described in Khresmoi Deliverable D1.4.2, "Report on Coupling Manual and Automatic Annotation" [16]. The reference corpus is a collection of Khresmoi web documents with entities annotated in the text, as described in Khresmoi Deliverable D1.4.1, "Khresmoi Manually Annotated Reference Corpus" [12]. Automatic annotations created at each iteration of the pipeline development process were corrected by multiple human annotators, and the differences between these manual annotations resolved to give the final corpus. Manual annotations were corrected according to a set of guidelines, described in Khresmoi deliverable D1.1 "Manual Annotation Guidelines and Management Protocol" [9], which were themselves based on the project requirements, described in Khresmoi Deliverable D8.2 "Use case definition including concrete data requirements" [17].

The results of comparing automatic annotations to the manual corrections are given in Khresmoi Deliverable D1.4.2, "Report on Coupling Manual and Automatic Annotation" [16]. The best results achieved are given in Table 2 below, achieved with a combination of rule based entity recognition, and a perceptron classifier trained on previous manual corrections.

		Strict			Lenient		
Key	**Response**	**P**	**R**	**F1**	**P**	**R**	**F1**
Manual corrections	Automatic annotations created by perceptron based annotation pipeline	0.90	0.65	0.76	0.98	0.70	0.82

Table 2: Performance of the Khresmoi annotation system compared to manual corrections

2.4 A comparison with external resources

The previous section outlined the internal Khresmoi evaluation of semantic annotation quality, against manual corrections in the Khresmoi reference standard. It is also useful to compare the Khresmoi system to other resources external to Khresmoi in order to place the Khresmoi semantic annotations in the landscape of annotation systems; to provide a check on annotation quality relative to those systems; and to learn lessons from those other systems. However, in doing such a comparison we must bear in mind that external resources may have been defined for a different end use, use different definitions of types, and be targeted at different text types. This means that there will be significant complexity in mapping the types and representation of one corpus to another, and a parallel complexity in interpreting results.

This section provides a comparison. The next sub-section describes corpora that were examined, and discusses their suitability for comparison to the Khresmoi system. This is followed by selection of a corpus, the CALBC Silver Standard [19], a description of its comparison to Khresmoi annotations, and the results of that comparison. The final subsection gives an analysis of these results.

2.4.1 Resources considered

Several resources containing standard annotations were considered for comparison to annotations produced by the Khresmoi pipeline. The criteria used to select a corpus for comparison were:

- Type of text. Standard corpora in the biomedical domain usually concentrate on publications and abstracts, although there are some medical record based corpora. Although Khresmoi does consider abstracts and radiology reports, the bulk of the application and its tuning, was directed at health information web pages.
- Type of annotations. Biomedical corpora annotate a variety of semantic types, most frequently diseases, genes and proteins. Khresmoi semantic types are taken from a broader set: diseases, investigations, anatomy, and drugs.
- Standardisation of annotations. The most accurate corpus against which we could compare the Khresmoi system would be a methodically constructed corpus of manual annotations. The best of these would involve a reconciliation of the work of multiple annotators, in order to reduce annotator bias. Such corpora are expensive, and other less accurate approaches are often taken in constructing corpora.

The resources considered are briefly described below, together with the reason for their rejection or use.

- MuchMore[1]. The MuchMore corpus is a corpus of medical journal abstracts, annotated with basic lexical and syntactic information, as well as UMLS semantic classes. The corpus contains parallel English and German texts, the primary use case being machine translation (MT). Although it does contain semantic classes relevant to Khresmoi, the corpus has been built by a single automatic system, and so is inevitably biased to that system.
- BioWSD[2]. BioWSD is a project on word sense disambiguation. The BioWSD corpus contains journal abstracts annotated with semantic classes. These are restricted in the case of one corpus to gene names and in the other corpus to multiple classes across abbreviations only. The corpora therefore do not contain anything comparable to Khresmoi semantic annotations.
- I2B2 corpora[3]. The I2B2 community annotation challenges have been run annually since 2006. In each year, entrants to the challenge have attempted to annotate clinical documents with a given type or class. These have generally been quite specific (e.g. smoking, obesity, temporal relations), rather than general types, and have been directed at clinical text. They therefore do not provide a good comparison for Khresmoi semantic annotations.
- BioCreAtIvE corpora[4]. The BioCreAtIvE corpora were also created for community challenges. They consist of journal text, annotated with proteins, genes and their interactions. This does not correspond to the types used in Khresmoi.
- CALBC Gold Standard[5]. The CALBC project (Collaborative Annotation of a Large Biomedical Corpus) created several corpora of biomedical abstracts annotated with various entities. The gold standard, while it does contain types that overlap with those used in Khresmoi, contains no CUI information, and would therefore not provide a full comparison with the Khresmoi semantic annotations.
- CALBC Silver Standard corpus[6]. The CALBC project also created silver standard corpora (SSC), containing annotations created automatically by several project participants, and harmonised across those participants. The corpus consists of biomedical journal abstracts, and annotations have some overlap with the Khresmoi annotation types. The corpus contains both TUIs and CUIs for annotations. This corpus was selected for further comparison.

2.4.2 Results of the comparison with the CALBC Silver Standards

The CALBC corpora [19] are annotated with four different semantic groups: Proteins and Genes (PRGE), Chemicals (CHED), Diseases and disorders (DISO), Living Beings (LIVB). These are defined in terms of TUIs, described in a set of guidelines. There are several corpora. For the comparison, the "CALBC-SSC-III-Small" was chosen. This was annotated by 13 different systems in the CALBC project, after which annotation boundaries were harmonised and annotations selected where there was agreement between 5 or more systems. The TUI and CUI of all annotations with such agreement are recorded for harmonised boundaries. One span of text may have, and usually does have, multiple TUI / CUI annotations. The corpus is thus a harmonised set of the commonalities between systems, the logic being that if many automatic systems agree, there is a high chance the annotation is correct.

1 http://muchmore.dfki.de/resources1.htm

2 http://nlp.shef.ac.uk/BioWSD/

3 https://www.i2b2.org/

4 http://biocreative.sourceforge.net/index.html

5 http://www.ebi.ac.uk/Rebholz-srv/CALBC/corpora/resources.html

6 http://www.ebi.ac.uk/Rebholz-srv/CALBC/corpora/resources.html

The size of the SSC-III-Small corpus, in numbers of documents, is given in Table 3. On processing, it was found that approximately one third of the documents in the corpus had no abstract text, and these were excluded from the comparison. The Khresmoi annotation pipeline does not create annotations similar to the CALBC PRGE, CHED and LIVB groups, and so these CALBC annotations were excluded from the comparison. Similarly, CALBC does not include groups similar to the Khresmoi Anatomy, Investigation and Drug semantic types, and so these could not be compared. This left the single CALBC group of DISO for comparison, which overlaps in its definition with the Khresmoi Disease type. The number of DISO annotations, broken down by TUI, is given in Table 4. The overlap with the Khresmoi Disease type can be seen in Tables 1 and 4, where CALBC DISO and Khresmoi Disease are defined in terms of UMLS TUIs. It should be noted that while the formal definition of CALBC DISO does not include TUI T020, DISO annotations marked with T020 are to be found in SSC-III-Small.

Documents in the corpus	174 999
Documents skipped because the abstract was missing	56 656
Documents remaining for comparison	118 343

Table 3: CALBC corpus SSC-III-Small, number of documents processed in the comparison with the Khresmoi annotation system

TUI	**TUI description**	**Number of annotations**
T019	Congenital Abnormality	10 144
T020	Acquired Abnormality	5 547
T047	Disease or Syndrome	1 561 585
T048	Mental or Behavioral Dysfunction	26 659
T050	Experimental Model of Disease	14 203
T190	Anatomical Abnormality	4 859
T191	Neoplastic Process	599 185

Table 4: Number of disoder (DISO) annotations found in SSC-III-Small. Multiple TUIs at a single location are each counted as one annotation.

The Khresmoi semantic annotation pipeline was run over the full set of abstracts from SSC-III-Small. The Number of Khresmoi annotations found by type are given in Table 5. The numbers of Disease annotations broken down by TUI are given in Table 6.

Khresmoi annotation type	Number
Disease	926 551
Anatomy	341 645
Investigation	453 464
Drug	641 254

Table 5: Semantic annotations created on SSC-III-Small by the Khresmoi annotation pipeline

TUI	TUI description	Number
T033	Finding	328 393
T037	Injury or Poisoning	13 372
T046	Pathologic Function	95 238
T047	Disease or Syndrome	323 236
T048	Mental or Behavioral Dysfunction	6 663
T049	Cell or Molecular Dysfunction	25 522
T050	Experimental Model of Disease	5 303
T184	Sign or Symptom	36 374
T191	Neoplastic Process	92 450
	All	**926 551**

Table 6: Khresmoi Disease annotations created on SSC-III-Small by the Khresmoi annotation pipeline, broken down by TUI

The Disease annotations shown in Table 6 were compared to the DISO annotations in SSC-III-Small. Several different comparisons were carried out, varying the annotations compared. These variations are described in the following bullet points, which correspond to columns and rows in the comparison results tables (Tables 7 and 8, to be discussed).

- **Response.** The response annotations, always the Khresmoi Disease annotations.

- **Key.** The key annotations. This was always the CALBC DISO annotations, restricted to those UMLS semantic types found in the Khresmoi Disease annotation definition. Two different key sets were used:
  - **Most frequent TUI.** The key consists of the DISO annotation defined by the single most frequent TUI at a span, selected from all of the DISO annotations at that span. In other words, we are comparing the TUI or CUI selected by the majority of CALBC annotation systems with the Khresmoi annotation.
  - **Any matching TUI.** The key consists of all of the DISO annotations at a span where they match a TUI from the set of TUIs defining Khresmoi Disease (see Table 1). In other words, we are comparing any CALBC annotation that could potentially be found by the Khresmoi system, to those that were actually found by the Khresmoi system.
- **TUI.** The key set is restricted to annotations for particular TUIs, as follows, allowing us to compare systems across specific TUI subsets:
  - **All.** All key annotations are considered, regardless of TUI.
  - **Individual TUI code.** Only those annotations matching the given TUI are included in the key set.
- **Agreement on.** A key / response pair is considered to be a match if there is agreement between:
  - **TUI.** The TUI in key and response annotation agree, giving a comparison on the broad TUI types.
  - **CUI.** The CUI in key and response annotation agree, giving a comparison on the more fine grained CUIs.

Given the size of the SSC-III-Small corpus and the number of comparisons we wished to make, we ran comparisons over a random sample of 2000 abstracts. The full set of results is shown in Table 7, and the results are analysed in the following section. In order to test the validity of the sampling used, we carried out two of the experiments across the full SSC-III-Small corpus. The results of this are shown in Table 8. The two rows correspond to the results for the 2000 abstract samples in rows 1 and 11 of Table 7. Results are similar for the full corpus and the sample, suggesting that the random sampling was valid.

			Strict			Lenient		
Key	TUI	Agreement on	P	R	IAA	P	R	IAA
Most frequent TUI	All	CUI	0.31	0.26	0.28	0.59	0.48	0.53
		TUI	0.38	0.31	0.34	0.73	0.60	0.66
	T047	CUI	0.30	0.26	0.28	0.58	0.50	0.54
		TUI	0.37	0.32	0.34	0.70	0.61	0.65
	T048	CUI	0.31	0.29	0.30	0.44	0.41	0.43
		TUI	0.32	0.30	0.31	0.49	0.45	0.47
	T050	CUI	0.07	0.10	0.09	0.46	0.64	0.53

		TUI	0.07	0.10	0.09	0.46	0.64	0.53
	T191	CUI	0.35	0.23	0.28	0.66	0.44	0.53
		TUI	0.45	0.30	0.36	0.87	0.57	0.69
Any matching TUI	All	CUI	0.35	0.28	0.31	0.67	0.54	0.60
		TUI	0.38	0.31	0.34	0.74	0.60	0.67
	T047	CUI	0.33	0.45	0.38	0.63	0.86	0.73
		TUI	0.33	0.45	0.38	0.63	0.86	0.73
	T048	CUI	0.33	0.53	0.41	0.54	0.85	0.66
		TUI	0.33	0.53	0.41	0.54	0.85	0.66
	T050	CUI	0.07	0.11	0.09	0.49	0.73	0.59
		TUI	0.07	0.11	0.09	0.49	0.73	0.59
	T191	CUI	0.40	0.43	0.42	0.79	0.84	0.81
		TUI	0.40	0.43	0.42	0.79	0.84	0.81

Table 7: Comparison of CALBC DISO and Khresmoi Disease annotations on a random sample of 2000 abstracts from the SSC-III-Small corpus, broken down by TUI. See the text for an explanation.

			Strict			Lenient		
Key	TUI	Agreement on	P	R	IAA	P	R	IAA
Most frequent TUI	All	CUI	0.33	0.27	0.29	0.60	0.49	0.54
Any matching TUI	All	CUI	0.37	0.30	0.33	0.68	0.55	0.61

Table 8: Comparison of CALBC DISO and Khresmoi Disease annotations on the full SSC-III-Small corpus. See the text for an explanation.

2.4.3 Analysis of results

In the below analysis, the span of an annotation is shown by square brackets. Thus, “premature [ovarian] failure” means that the discussion is referring to an annotation that spans the word “ovarian”. Sometime, CUIs and TUIs will also be shown in the square brackets, so that "[carcinoma,C0007099/T191]” means that “carcinoma” has been annotated, with a CUI of C0007099 and a TUI of T191.

Three underlying trends are apparent in the results of Table 7:

- Agreement on TUI is better than agreement on CUI, where the key is the most frequent TUI. This is to be expected: the task of finding a broad semantic group is easier than that of finding a specific class.
- Agreement is higher where the key set contained any matching TUI, as opposed to the key set based on the most frequent CALBC TUI. This is because on occasions, the most frequent CALBC TUI is one that could not possibly be chosen by the Khresmoi system, as it is not part of the definition of the Khresmoi disease annotation type.
- There is a large difference between strict and lenient scores, across all comparisons. This is to be expected: the task of finding not only the correct semantic type but also the correct span is harder. It should be noted, however, that SSC-III-Small is itself based on harmonised annotation spans, and so a comparison using the lenient score is more meaningful: we have not considered the annotation spans given by the multiple individual systems, and this information is not available at the per-system level in SSC-III-Small.

In general, the Khresmoi annotations appear to compare badly with the CALBC Silver Standard. As the CALBC Silver Standard is a thought through and methodologically created standard, this would suggest that the Khresmoi annotations are of low quality. This does not, however, agree with the results from manual correction of the Khresmoi annotations. The comparison with CALBC therefore requires further analysis.

First, we must consider that SSC-III-Small is a *silver* standard, so called because it is created from the agreement between multiple systems. In order to interpret the level of Khresmoi agreement with SSC-III-Small, we need to look at the underlying agreement between the CALBC constituent systems. This is given in [21], for DISO, as F measure between 0.58 to 0.81,
with an average of 0.71. It is not clear how this F measure is derived, but given the harmonised nature of the SSC-III-Small, this must be a lenient measure, and therefore comparable to the lenient measures in the “any matching TUI” rows of Table 7. In this light, the Khresmoi system compares favourably to the CALBC systems.

Second, we must consider that the set of TUIs found by the CALBC systems are not the same as those found by the Khresmoi system. This was allowed for in the comparison, by only comparing the Khresmoi system with the TUIs it could find in SSC-III-Small. The Khresmoi system, however, may still show degraded agreement scores. This is apparent in a document-by-document analysis of differences between Khresmoi and CALBC. The largest difference found was due to the difference between the TUIs used to define CALBC DISO and Khresmoi Disease. In particular, Khresmoi includes TUI T033 (Finding) in the definition of Disease. Terms in UMLS that are typed as T033 are often ambiguous with a disease of the same name, i.e. type T048 (Disease). As a result of this, some text is annotated by the Khresmoi system as T033 where it is annotated as T048 by CALBC. This therefore counts as a missed annotation in the comparison. Which of the CALBC annotation or the Khresmoi annotation is correct would have to be decided on a case by case basis.

The difference in definition of Disease / DISO is not restricted to Khresmoi and CALBC, and is reflected in differences with the similar DISO semantic group in to which UMLS groups its own TUIs [20]. The TUIs comprising the UMLS DISO group is shown in Table 9, alongside those comprising CALBC DISO and Khresmoi Disease. It can be seen that Khresmoi is in fact closer to the UMLS definition of DISO than CALBC, differing only in three anatomical abnormality TUIs, which Khresmoi includes instead in its Anatomy group.

		UMLS DISO Semantic Group	CALBC DISO annotation	Khresmoi Disease type
T019	Congenital Abnormality	Y	Y	
T020	Acquired Abnormality	Y	Y	
T033	Finding	Y		Y
T037	Injury or Poisoning	Y		Y
T046	Pathologic Function	Y		Y
T047	Disease or Syndrome	Y	Y	Y
T048	Mental or Behavioral Dysfunction	Y	Y	Y
T049	Cell or Molecular Dysfunction	Y		Y
T050	Experimental Model of Disease	Y	Y	Y
T184	Sign or Symptom	Y		Y
T190	Anatomical Abnormality	Y	Y	
T191	Neoplastic Process	Y	Y	Y

Table 9: A comparison of the TUIs in the UMLS and CALBC DISO semantic groups, and the Khresmoi Disease semantic type

The third important insight is that CALBC splits up multi-word terms and sometimes assigns the same CUI to every word in a term, giving multiple targets where there should not be any. In other cases, CALBC assigns CUIs to some but not all of the words. This penalises Khresmoi in the strict measures, but to also to some extent the lenient measures.

Very often, the CALBC corpus does not annotate the whole span of a multi-word term. For example there are several occurrences of "Addison's disease" or "Autoimmune Addison's disease". All of them have each individual word annotated, “[Autoimmune] [Addison]'s [disease]”, and each is assigned the same CUI, C0271737. The Khresmoi system would annotate the whole string with the same CUI. Because of the additional number of targets, the agreement between systems drops.

Another example is given by: “squamous epithelial dysplasia and squamous cell carcinoma of the esophagus”. CALBC annotates this as: "squamous epithelial [dysplasia] and [squamous] cell [carcinoma] of the [esophagus]" assigning CUIs to [dysplasia], [squamous],[carcinoma], and [esophagus]. Khresmoi annotated [epithelial dysplasia], and [squamous cell carcinoma] as a finding.

In the same document, "carcinoma in situ" is annotated by CALBC as "[carcinoma,C0007099/T191] in [situ,C0007099/T191]” and by Khresmoi as "[carcinoma in situ, C0007099/T191”] The multiple targets in the CALBC key mean that, all of recall, precision and IAA are degraded.

Similar examples are given by what appears to be a CALBC bias towards shorter and less specific annotations. This is possibly due to the CALBC annotation harmonisation process. For example: "human immunodeficiancy virus" is annotated fully and with the correct CUI (C0019693) by the Khresmoi system, but only "immunodeficiancy" (C0021051) is annotated in SSC-III-Small.

The case of partial annotation of terms may be combined with Khresmoi's annotation using a non-CALBC TUI. For example, "premature ovarian failure" is annotated in the SSC-III-Small "premature [ovarian] failure" while Khresmoi annotates "[premature ovarian failure]" as a finding.

All these examples help to explain the large difference between strict and lenient measures and seem to hint that the lenient measures are actually not far from the truth. IAA would be higher still if CALBC merged the separate annotations with same CUI.

The fourth and final point is that there are cases where one of either CALBC or Khresmoi is clearly wrong. For example: "sterility" gets annotated as C0021359/T047 by CALBC (disease in humans), even when discussing sterile yeasts, which should be annotated as C0678108/T046 (pathological function), as it is by Khresmoi. However, there are also cases when the Khresmoi system is wrong. A particular set of examples is triggered by Khresmoi rule-based recognition based on part-of-speech patterns. So if there is a dictionary match and the POS-tag pattern fits, the Khresmoi system will always assign something. This is sometimes wrong (e.g. because of the POS tagger being wrong) and it is often wrong for abbreviations. Another group of spurious Khresmoi annotations is based on a combination of not detecting actual multi-word entities and instead annotating some part of the term.

2.5 Conclusion of this chapter

The Khresmoi manual annotation task has allowed us to measure the quality of Khresmoi semantic annotation against manual corrections of those annotations. This shows a high level of agreement.

In order to validate Khresmoi semantic annotation in the wider context of biomedical annotation, we provided a comparison with other systems. Khresmoi annotates a wider set of semantic types than most biomedical annotation systems, and is unique in its primary focus on health information web pages. This makes a formal, quantitative comparison with other systems difficult. There is, however, significant work on disorder and disease annotation, and we have therefore compared Khresmoi annotations to diseases in the CALBC Silver Corpus. Khresmoi compares well, having a similar agreement to that between other CALBC source systems. An error analysis shows that, as might be expected, Khresmoi and CALBC take different approaches to the definition of annotation types and spans, and so any comparison must be treated with caution. The comparison provides an interesting insight in to the difficulty of such comparisons, and useful information about the Khresmoi system, the standards to which it is being compared, and their relative position in the biomedical annotation landscape.

3 K4E[1]: automatic trustability detection evaluation and semantic search implementation

3.1 Automatic trustability detection

A system for automatic detection of the web page conformance to any of the 8 HONcode principles was developed at HON. Excerpts extracted by experts in the manual accreditation process, justifying the respect of the given criteria, were used as a training/test collection for this system. Details about this system are given in deliverable 1.6 [2]. The same document gives the effectiveness evaluation of the developed system for the English language.

HON has extended the system described in [2] to cover an additional set of languages namely: French, German, Spanish, Italian and Dutch. Table 1 gives the size of the corpus available for each criterion, for all the languages covered.

[1] KhresmoiForEveryone

It can be seen from this table that the size of French corpora is comparable to that of English. On the other hand, for other languages the corpora size is significantly smaller, with only 15 documents for the "Advertising policy" in Dutch.

This system is integrated into the K4E information extraction pipeline described in [3]. The results returned by this system are stored in the NoSQL database, and made available to project partners and implemented within the K4E search engine.

3.1.1 Evaluation

The deliverable [2] gives the evaluation of the system for automatic detection of HON code principles for the English. In this section we present the results of the evaluation for other languages.

The following set of parameters is chosen for this comparative evaluation:

- Learning algorithm: Naive Bayes
- Feature type: W1 (single word)
- Feature selection: TF-IDF
- Percentage of features kept: 30%

	English	*French*	*Spanish*	*German*	*Italian*	*Dutch*
Authority	2812	2338	894	493	500	107
Complementarity	2835	2005	819	567	490	106
Privacy	2683	2055	850	470	33	100
Reference	2349	1888	707	405	358	96
Justifiability	872	827	310	35	166	37
Transparency	2861	2349	88	604	414	121
Financial disclosure	2700	2098	814	546	489	103
Advertising policy	1412	627	433	246	255	15
Date	2794	2158	862	570	505	109

Table 10: Number of extracts per language for each criterion

We present the standard measures namely precision(P), recall(R) and F1-measure [4].

The results show similar tendencies for English and French. The "Reference" and "Justifiability" remain the "difficult" criteria to detect for these languages as well.

The results also illustrate the problems which might appear when the collection used for learning/test is too small. For certain criteria we are facing clear over fitting due to restrained vocabulary within the documents of the collection. Even though the obtained scores are quite high in terms of both recall and precision (e.g. "Transparency" for Spanish), they need to be verified in the manual vs. automatic evaluation. On the other hand, for Dutch, the system is unable to perform classification, because of too small of a number of documents (e.g. "Justifiability" 37, "Advertising policy" 15).

3.1.2 Search Engine Integration

	English			*French*			*Spanish*		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Authority	0.61	0.70	0.66	0.68	0.79	0.73	0.79	0.44	0.56
Complementarity	0.79	0.96	0.87	0.93	0.94	0.94	0.90	0.90	0.89
Privacy	0.95	0.97	0.96	0.93	0.98	0.96	0.99	0.87	0.92
Reference	0.59	0.58	0.59	0.73	0.43	0.54	0.77	0.35	0.49
Justifiability	0.65	0.25	0.36	0.83	0.43	0.57	1.00	0.04	0.07
Transparency	0.93	0.93	0.93	0.93	0.90	0.91	1.00	0.73	0.85
Financial disclosure	0.70	0.72	0.71	0.80	0.88	0.83	0.79	0.52	0.63
Advertising policy	0.74	0.73	0.74	0.92	0.37	0.53	0.86	0.14	0.25
Date	0.96	0.95	0.95	0.98	0.92	0.95	0.99	0.91	0.95

Table 11: Trustability evaluation results for English, French and Spanish

	German			*Italian*			*Dutch*		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Authority	0.69	0.77	0.73	0.81	0.74	0.77	0.60	0.33	0.43
Complementarity	0.95	0.98	0.96	0.90	1.00	0.95	1.00	0.60	0.75
Privacy	0.98	1.00	0.99	0.91	0.98	0.94	1.00	0.30	0.46
Reference	0.52	0.58	0.55	0.75	0.82	0.78	0.50	0.11	0.18
Justifiability	0.00	0.00	0.00	1.00	0.44	0.61	0.00	0.00	0.00
Transparency	0.93	0.90	0.91	0.94	0.85	0.89	1.00	0.54	0.71
Financial disclosure	0.79	0.83	0.81	0.80	0.85	0.85	1.00	0.20	0.33
Advertising policy	0.77	0.42	0.54	0.73	0.50	0.59	0.00	0.00	0.00
Date	1.00	0.96	0.98	0.98	0.96	0.97	1.00	0.70	0.82

Table 12: Trustability evaluation results for German, Italian and Dutch

For each result returned by the K4E search engine the information concerning the level of trust automatically detected is displayed (Illustration 1).

It is available under advanced search results of the K4E. Trustability level is displayed in the form of the red-orange-green icon. It is calculated over all crawled pages from the given host. Hovering over this icon, the user gets the information on the level of trust detected. In cases where the level is less than 100% missing criteria will be displayed as well (Illustration 2).

Illustrations 3-7 show the implementation of the trustability into the K4E for French, Spanish, German, Italian and Dutch respectively.

3.2 K4E: Semantic search implementation in Czech

K4E semantic search interface is available in all European languages. Apart from the semantic search available in English and described in [5], the semantic search is now available in Czech. In this language the user is given the possibility to choose from the list of subjects given in the Illustration 8 such as: “Nemoc nebo syndrom”, choose from the list of corresponding predicates (Illustration 9). In the choice of search object the user has the possibility to either choose from the proposed list (Illustration 10) or to type in the object.

Illustration 12 gives the results returned for the query “"Nemoc nebo syndrom' 'lecen/-a/-o' 'Lasix"” (“Disease or syndrome treated by Lasix”).

4 Enhanced Textual Search Using Collaborative Filtering Methods

Retrieving the required information in response to users’ queries is a key requirement of the Khresmoi system. Various information retrieval (IR) approaches have been used to improve the effectiveness of search results. In this section, we explore the method of employing recommender systems to aid the standard IR system to obtain better retrieval results based on previous users’ search behaviours.

4.1 Method Framework

We use collaborative filtering approaches to aid the standard IR system. We believe that considering the previous users’ search behaviours to form a combined results list can improve the effectiveness of the IR retrieval model. Figure 1 shows the basic framework of this integrated model [8].

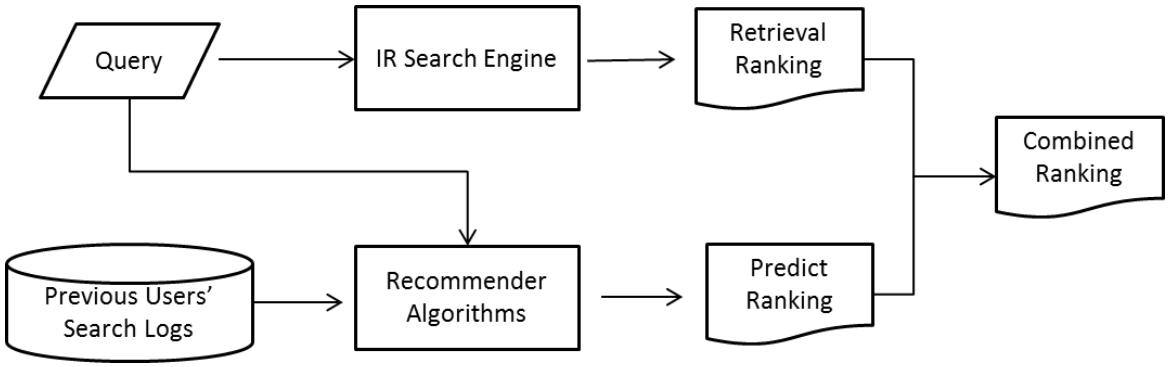

Figure 1: Basic framework of the integrated model.

In this integrated model, the user's query is given to the IR search engine to generate a ranked result list. The query is also passed to the recommender component. Here the similarity is computed between query representation and previous users' search logs. Users with similar preferences are found. Their logs are then used to generate recommendations for the current user's query. Finally, the IR results list and recommender prediction results list are combined and used as the final search output provided to the user.

4.2 Data Set

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.

From previous users' search logs, stored in the 'ula_logevent' table in ezDL, the event pairs with name label ('dlobjectqueryask', 'dlobjectquerytell') denote the user's query and his/her click-through information for this query. Based on this, the query and search activities are extracted from the user logs documents.

Query: Events with 'dlobjectqueryask' name label are derived from the 'ula_logevent'. They are the queries entered by the users. 10713 queries are extracted in total. Since we only focus on the textual content, we remove both the image query and the meaningless queries, such as 'test' '1'. In the end, 7161 text queries are kept. 6000 of them are selected randomly as the training set, and the other 1161 queries are used as test set.

Click-through information: Events with the name 'dlobjectquerytell' are the user's search activity. This click-through information is extracted from the 'ula_logeventparams' table in ezDL. Each user has a sequence of clicked pages. Figure 2 shows an example of a user's click-through information. Again, we only consider the textual content, the image pages are removed from each user's search activities. This results in a set of 3374 URLs. The number of clicked pages is far less than the number of queries because queries are duplicated. There are a large number of user queries on the same topic, e.g. 'diabetes' has been queried 665 times and 'heart' 521 times. For these duplicate queries, their clicked-through search logs contain large number of the same pages. This duplication is good when recommendations for the current user with an interest in a similar topic are being made.

eventid	paramname	paramvalue	sequence
12	item	mimir:knowyourabcs:2011:h:qes.subject.all	0
12	item	term:C0850356	1
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00241956:The Royal Melbourne Hospital	2
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00250731:State University of New York Upstate Medical University	3
12	item	term:C0085207	4
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00201292:Department of Health-Yuli Hospital	5
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00245882:VA Pittsburgh Healthcare System	6
12	item	term:C0011848	7
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Center for Neurologic Research	8
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:University of New Mexico	9
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Hartford	10
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Denver Health and Hospital Authority	11
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Beth Israel Deaconess	12
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:George Washington Univeristy	13
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Via Christi Regional Medical Center	14
12	item	clinicaltrial:http://data.linkedct.org/resource/trials/NCT00091949:Yale University	15

Figure 2: Example of a user's click-through information.

Data Collection:

Because the recorded click-through information is the URLs of the pages, we need to analyse the content for each page to generate the retrieval output and recommendation results. We crawled[1] the content of all users' clicked pages using Scrapy[2]. The textual content extracted for each page constitutes the data collection we used for this experiment.

Relevance assessment:

For each obtained query, its corresponding list of clicked pages is used as the qrel file (relevance assessment) for the query.

4.3 Content-Based Filtering Algorithm

Content-based filtering algorithms are based on the content of the document and a profile of the current user. In a content-based recommender system, keywords are used to present the document. Also a user profile needs to be built to indicate the type of document a user likes. In other words, these algorithms are trying to provide users with the documents which are similar to documents they liked in the past. Different retrieval models are used in content-based filtering algorithms, such as keyword matching or the Vector Space Model (VSM) with basic TF-IDF (Term Frequency-Inverse Document Frequency) weighting. In this experiment, we choose the most common term weighting scheme, TF-IDF weighting, which is based on empirical observations regarding the nature of text (Salton, 1989):

- Rare terms are more valuable than frequent terms (IDF assumption) across a collection of documents.
- Multiple occurrences of a term in an individual document are more valuable than single occurrences (TF assumption)
- Long documents are not preferred to short documents (normalization assumption)

In other words, terms that occur frequently in one document but rarely in the rest of a corpus, are more likely to be significant in describing the topic of the document. In addition, normalizing the resulting weight vectors prevents longer documents from having a greater chance of retrieval at high rank. These assumptions are exemplified by the TF-IDF function shown in Equation (1).

$$TF\text{-}IDF(t,d) = TF(t,d) \cdot log\frac{N}{n_t} \qquad (1)$$

Where $TF(t,d)$ is the term frequency of term t in document d. N denotes the number of documents in the corpus, and n_t denotes the number of documents in the corpus in which the term t occurs at least once.

In order for the weights to fall in the [0,1] interval[3] and for the documents to be represented by vectors of equal length, weights obtained by Equation (1) are usually normalized using the cosine normalization shown in Equation (3).

[1] The collected pages URLs have been changed; please check Appendix B for the url change rules.

[2] http://scrapy.org/

[3] Here we do the normalizing for the term weight because we need to compare the document vector with the user profile vector whichs term weight also fall into [0,1] interval.

$$Tw_{t,d} = \frac{TF\text{-}IDF(t,d)}{\sqrt{\sum_{s=1}^{|T|} TF\text{-}IDF(t_s,d)^2}} \quad (2)$$

where t_s indicates a term occurs in document d. And $|T|$ is the number of all terms in document d.

A similarity measure is required to determine the closeness between two documents. Many similarity measures have been derived to describe the proximity of two vectors; among those measures, cosine similarity shown in Equation (3) is the most widely used.

$$sim(d_i,d_j) = \frac{\sum_t w_{ti} \cdot w_{tj}}{\sqrt{\sum_t w_{ti}^2} \cdot \sqrt{\sum_k w_{tj}^2}} \quad (3)$$

In content-based recommender systems relying on the VSM, both user profiles and items are represented as weighted term vectors. Predictions of a user's interest in a particular page can be derived by computing the cosine similarity between them.

4.4 Experiment Set Up

IR component:

The Terrier BM25 retrieval model was used to generate ranked lists for the IR component. The BM25 probabilistic model is defined in [5], and is based on a body of prior work on probabilistic ranking. A Terrier system stopword list of 733 words was used with a Porter stemmer [6] to pre-process the input text.

Recommender component:

For the recommender component, the first step was to find the other users with similar interests, then to generate the recommendation for a query based on the click-through information from these users.

In this data collection, since many users share the same interests, we simply compare the current user's query with other queries, if it is the same or shares a term with other queries, we consider them similar.

Since rating information is not present in the recorded user logs, for this experiment we choose the content-based filtering (CBF) algorithm to generate prediction results for the integrated model. With the CBF method, the similarity between documents and the current query needs to be computed.

The length of the query is usually too short to compute the current query and documents similarity reliably, to address this problem we use the centroid representation [7] for the current query to perform the similarity computation.

Centroid Document:

In order to generate the centroid document for the current query q: First we obtain the initial retrieval list for this query. Then for each document d in this retrieval list, stopwords are first removed with subsequent application of Porter stemming. The resulting document vector was then weighted using TF-IDF to produce a weighted vector $d_{tf\text{-}idf} = (tf\text{-}idf_1, tf\text{-}idf_2, \dots)$, where $tf\text{-}idf$ is the frequency inverse document frequency of the i^{th} term.For the set N of documents and their corresponding vector representations, we define the centroid vector C_q for the query using Equation (4).

$$C_q = \frac{1}{|N|} \cdot \sum_{i=1}^{N} d_i \quad (4)$$

In order to generate the centroid documents for queries, the steps are:

- For a query, obtain its retrieval list.
- Take the top 3 items from the retrieval list to generate its centroid document. This top 3 is set based on the results of a training set.

The procedure for application of the CBF method to output recommendations for each test query is as follows:

- Generate the centroid document for the current query.
- For each query, compute the similarity between the representation of the query and each document using cosine similarity, and rank all documents in descending order based on their distance from the current query centroid document.

Combine results:

The CombANZ operator is used in this experiment to integrate a prediction rank into the retrieval list..

$CombANZ_i$ is specified as the same sum of document scores as $CombSUM_i$ but divided by the number of ranking schemes which contain document i.

$$CombANZ_i = \frac{CombSUM_i}{number\ of\ nonzero\ score_i} \quad (5)$$

4.5 Results

In the following result table, IR indicates the result for the standard search engine result without query expansion (QE). IR+QE represents the IR results with QE which was conducted using BM25 from Terrier. And IR+QE+CBF is the results obtained by combining the standard IR+QE results with the output obtained from the recommender (CBF).

	MAP	P@5	P@10	P@20
IR	0.4109	0.3726	0.3167	0.2730
IR+QE	0.4354	0.3973	0.3521	0.2789
IR+QE+CBF	0.5015	0.4289	0.3814	0.2936

Table 13: Results comparison for standard IR results with the output of the integrated model.

Looking at the results obtained, Table 13, we can clearly see that the integrated model outperforms the standard IR search engine.

For this experiment, we could see that solely using the IR system also obtains relativly good search results, this is because the queries in the log describe the users information needs well. Even though most queries only consist of one word, they are still indicative of users' information need. However,

our integrated method obtains better results. The reason for this is that one of the key aims of the collaborative filtering method is to detect novel items. In this experiment, we exploit the search logs of the users with similar interests to the current query to generate recommendations for the current user. These related users may view some pages which are relevant to the current user's information need which do not appear in the retrieval result. In this situation, novel pages can be detected and added to the final combined results to further improve the MAP of the results.

5 Conclusion

This deliverable describes the results of the second evaluation phase for Workpackage 1. Three sections are presented, two of them are the extension of previous work and one contains a newly proposed evaluation method.

We have presented an evaluation of the semantic annotation indexed in Khresmoi, as used to drive semantic search techniques in the Khresmoi prototypes. This evaluation was of the quality of the annotations, abstracted away from their use in the Khresmoi system. We summarised a manual evaluation of the annotations, based on a comparison of those annotations to a manual correction. We reported a high level of agreement between the annotations and their manual corrections. We also provided a comparison of the Khresmoi annotations to annotations created independently of Khresmoi, the CALBC standard. The comparison is favourable, although differences between Khresmoi and CALBC (and other standards), makes such a comparison fraught with difficulties. The comparison does, however, allow us to position the Khresmoi annotation pipeline in the wider landscape of such systems.

Also, following the previous text categorization task introduced in deliverable 1.6, HON introduce an automatic trustability detection method, and apply it on different European languages. The results show similar tendencies for English and French language.Also, beyond English, different languages have been implemented into the K4E search engine. And Czech is implemented into the K4E semantic search interface.

A different retrieval method is also introduced in this deliverable. Besides using an IR system solely to discover the results for users, in this work, an integrated model is proposed, which combines the IR system with a recommender component. A content-based collaborative filtering method is used to generate prediction results for the recommender component. Experimentation is conducted on a relatively small collection, and results show this integrated model has the potential to improve the effectiveness of retrieve results.

6 References

[1] Lenzerini, M. (2002): Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, New York, 233–246.

[2] Allan Hanbury, Célia Boyer, Ljiljana Dolamic and Joaõ Palotti (2013): D1.6: Report on automatic document categorization, trustability and readability, Khresmoi *project public deliverable.*

[3] Allan Hanbury, William Belle, Nolan Lawson, Ljiljana Dolamic, Natalia Pletneva, and Matthias Samwald. (2012): D8.3: Prototype of a first search system for intensive tests. *Khresmoi project public deliverable.*

[4] C. J. VAN RIJSBERGEN. *Information Retrieval.* Butterworths, London, UK, 1979.

[5] Robertson S.E., Walker S., and Jones M.S., "Okapi at TREC-3," *In the Proceeding of Second Text Retrieval Conference. (TREC-3)*, (1995)

[6] Porter M.F., An algorithm for suffix stripping, *In Book "Readings in Information Retrieval" page 313-316*. (1980)

[7] Radev D.R., Jing H.Y., Styś M. and Tam D., Centroid-Based Summarization of Multiple Documents. *In Proceeding of Emerging Trends in Engineering and Technology, ICETET'08*. (2008)

[8] Li W. and Jones G.J.F., Enhanced Information Retrieval by Exploiting Recommender Techniques in Cluster-Based Link Analysis. *In Proceeding of ICTIR'13*. (2013)

[9] Roberts, A., Aswani, N., Pletneva, N., Boyer, C., Heitz, T., Bontcheva K., Greenwood, M.A., D1.1 Manual Annotation Guidelines and Management Protocol, February 2012.

[10] Greenwood, M.A., Roberts A., Aswani, N., Gooch, P., D1.2 Initial prototype for semantic annotation of the Khresmoi literature, May, 2012.

[11] Momtchev, V., D5.2 Large Scale Biomedical Knowledge Server, May 2012.

[12] Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A., Building a semantically annotated corpus of clinical texts, Journal of Biomedical Informatics. 42 (2009) 950-966.

[13] Hripcsak, G, Rothschild, A., Agreement, F-measure and reliability in information retrieval. J Am Med Inform Assoc. 2005 May-June;12(3):296–298.

[14] Niraj Aswani, Liadh Kelly, Mark Greenwood, Angus Roberts, Matthias Samwald, Natalia Pletneva, Gareth Jones, Lorraine Goeuriot. Report on Results of the WP1 First Evaluation Phase, Khresmoi project deliverable D1.3, August 2012.

[15] Betsy L. Humphreys and Donald A.B. Lindberg and Harold M. Schoolman and G. Octo Barnett. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998, 5:1

[16] Angus Roberts, Johann Petrak, Niraj Aswani, Report on Coupling Manual and Automatic Annotation, Khresmoi project deliverable D1.4.2, May 2013

[17] Use case definition including concrete data requirements. Khresmoi project deliverable D8.2

[18] Mark A. Greenwood, Angus Roberts, Niraj Aswani, Johann Petrak. Manually Annotated Reference Corpus, Khresmoi project deliverable D1.4.1, May 2013.

[19] Rebholz-Schuhmann, D., A. Jimeno Yepes, E. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. (2010) "CALBC Silver Standard Corpus." J Bioinform Comput Biol. 2010 Feb;8(1):163-79.

[20] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Proceedings of Medinfo 2001;10(Pt 1):216-20.

[21] Rebholz-Schuhmann, Dietrich, Jimeno-Yepes, Antonio José, van Mulligen, Erik M., Kang, Ning, Kors, Jan, Milward, David, Corbett, Peter, and Hahn, UdoThe CALBC Silver Standard Corpus -
Harmonizing multiple semantic annotations in a large biomedical corpus. Proceedings of the 3rd International Symposium on Languages in Biology and Medicine 2009 Pages 64 – 72. November 2009.

Appendix A:

K4E: Trustability implementation

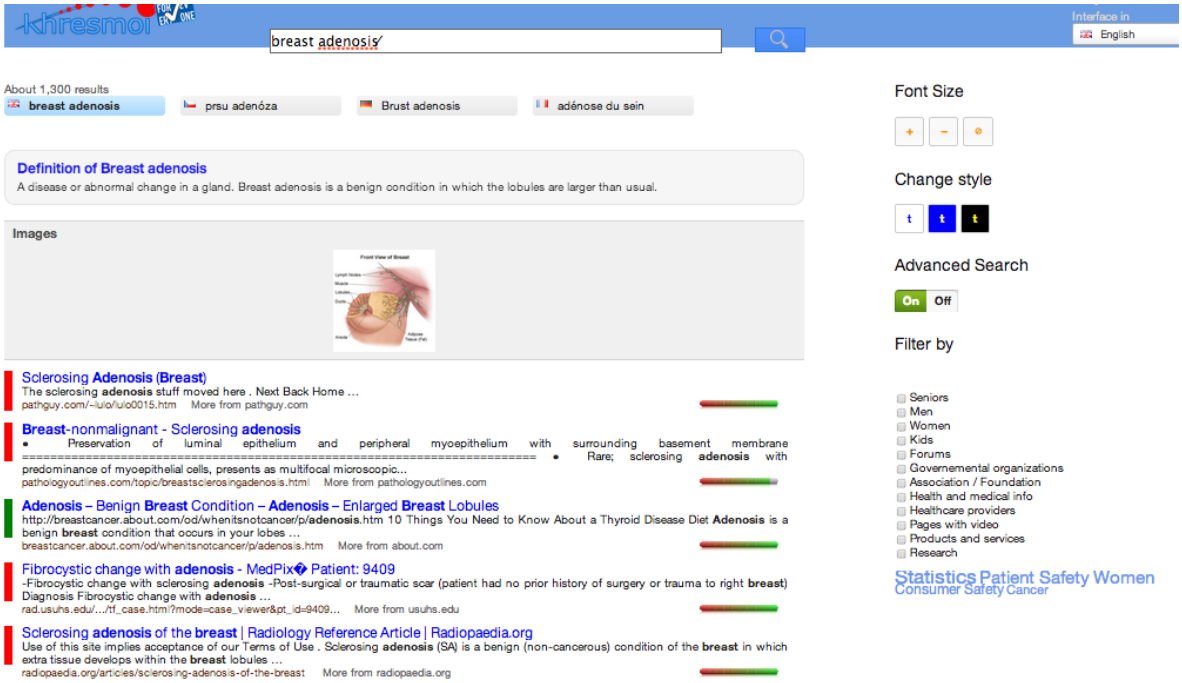

Illustration 1: Trustability implementation

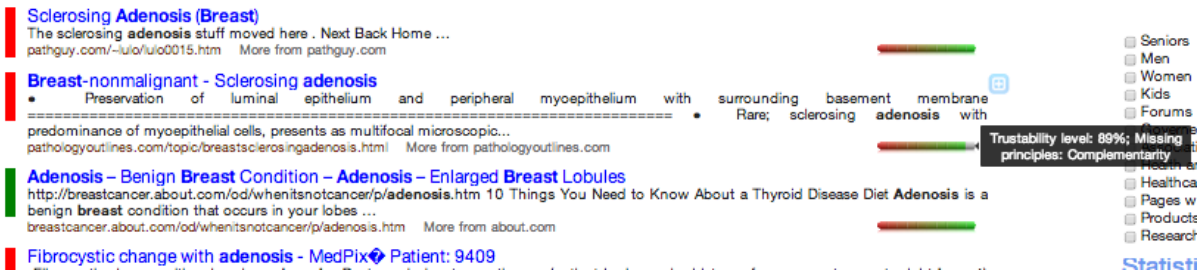

Illustration 2: Trustability, missing principles (English)

Illustration 3: Trustability French

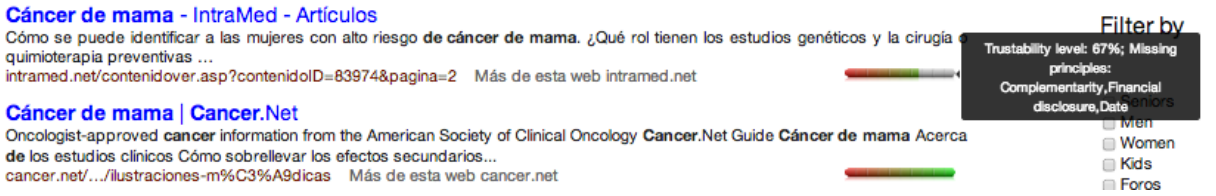

Illustration 4: Trustability Spanish

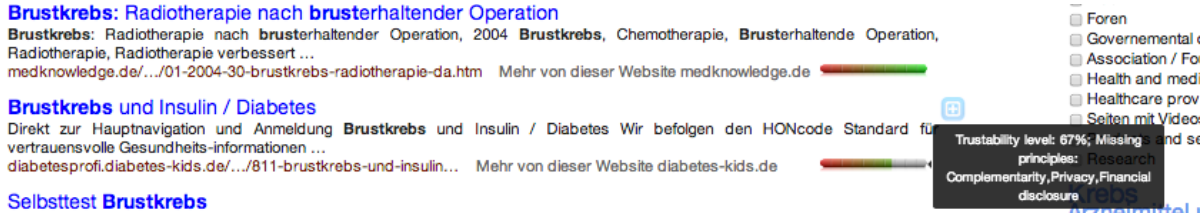

Illustration 5: Trustability German

Illustration 6: Trustability Italian

Illustration 7: Trustability Dutch

K4E: Semantic search implementation

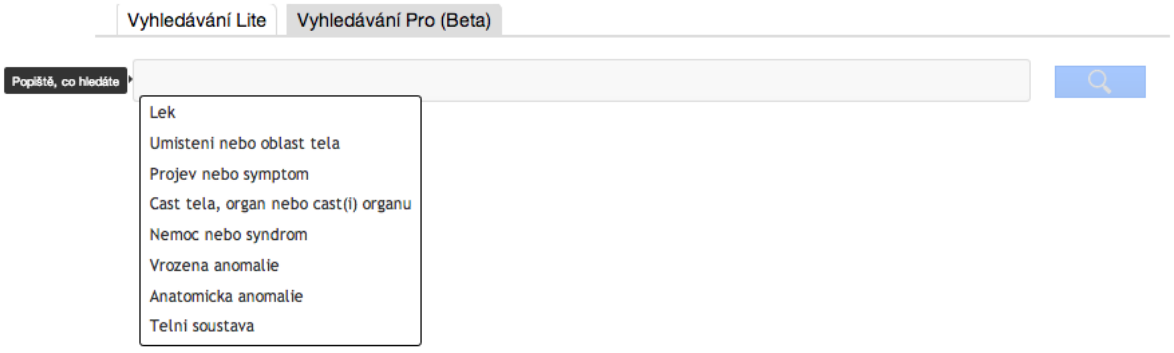

Illustration 8: Czech: Subject selection

Illustration 9: Czech: Predicate selection

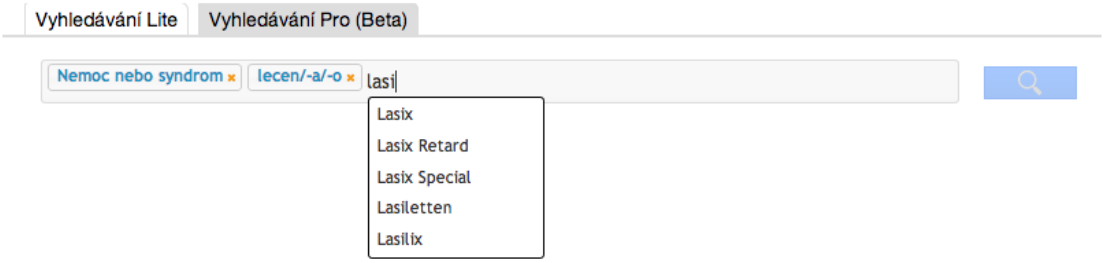

Illustration 10: Czech: Object selection

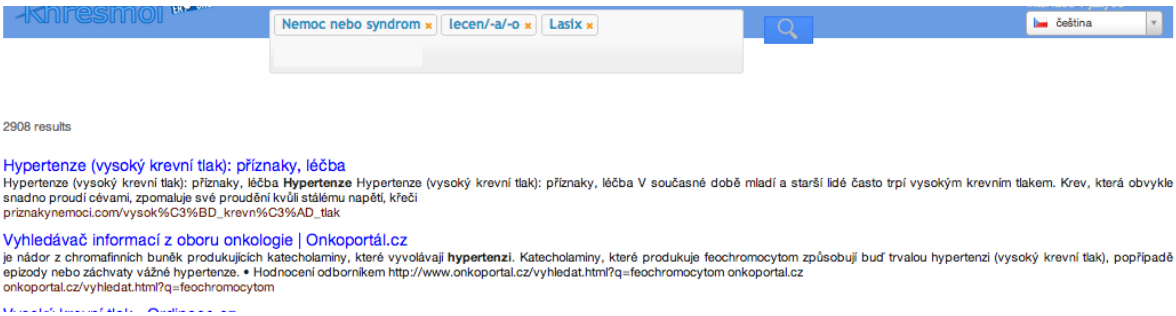

Illustration 11: Czech: search result page

Appendix B:

URL Change Rules

Clinical trial:http://data.linkedct.org/resource/trials/NCT00001458:National

Institutes of Health Clinical Center, 9000 Rockville Pike

--> a clinical trial used during an early stage of the project, they changed their directory structure slightly. In order to retrieve the content, the 's' from 'trials' has to be removed from the URL part

e.g. http://data.linkedct.org/resource/trial/NCT00001458/

drug:http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00624'

--> like the clinical trial, drugs were fetched during an early stage of the project; they directly contain the URL of the resource.

e.g.

http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00624 (which

goes to

http://wifo5-03.informatik.uni-mannheim.de/drugbank/page/drugs/DB00624)

term:http://linkedlifedata.com/resource/umls/id/C0011854'

--> same as clinical trials and drugs, the URLs should still resolve 'term:C0491562'

--> this was a shorter form of the same id type