Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Public version and updated report on the KHRESMOI data collection

Deliverable number	*D6.2*
Dissemination level	*Public*
Delivery date	*28 February 2013*
Status	*Final*
Author(s)	*Priscille Schneller, Khalid Choukri, Julien Gobeill, Liadh Kelly, Georg Langs, Dimitrios Markonis, Pavel Pecina, Konstantin Pentchev, Natalia Pletneva, Angus Roberts.*

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 Deliverable is a final report on the activities carried out within Work Package 6 of the Khresmoi project related to the identification, rights negotiation, crawling, and preparation of data resources (Task 6.1). It is an updated version of the Deliverable 6.1, the initial version and report on the Khresmoi data collection.

Khresmoi is developing a multilingual multimodal search and access system for biomedical information and documents. Several components of the search engine are based on the use of various Language Resources (LRs). In addition, a knowledge base is constructed to improve query performance.

The main purpose of this report is to describe the status of LRs identified, used and/or produced within the Khresmoi project. We elaborate on them in this deliverable along four main axes:

1) Existing LRs that have been used for training systems, evaluating the search engine components, or to be integrated into the knowledge base: can be text-centric datasets, image data or terminological resources such as ontologies.

2) LRs produced during the project for training or evaluation purposes: when existing datasets did not fit the needs, some LRs have been adapted or created for specific purposes.

3) Data sources indexed in the current prototype: websites identified to be relevant when looking for medical information.

4) Legal issues and exploitation rights: for the Khresmoi project consortium, it is crucial that the permission to use copyrighted LRs is obtained.

Table of Contents

List of Abbreviations

CUNI	**Charles University, Prague**
FMA	**Foundational Model of Anatomy**
GAW	**Society of Physicians in Vienna**
HON	**Health on the Net**
IE	**Information Extraction**
IR	**Information Retrieval**
KB	**Khresmoi Knowledge Base**
LR	**Language Resource**
PACS	**Picture Archiving and Communication Systems**
MT	**Machine Translation**
NLM	**National Library of Medicine**
TUW	**Technical University of Vienna**
UMLS	**Unified Medical Language System**
USFD	**The University of Sheffield**
WIPO	**World International Property Organization**

1 Introduction

This Deliverable is a final report on the activities carried out within Work Package 6 of the Khresmoi project related to the identification, rights negotiation, crawling, and preparation of data resources (Task 6.1). It is an updated version of the Deliverable 6.1, the initial version and report on the Khresmoi data collection. Unlike D6.1, this deliverable is a public report.

The Khresmoi main objective is to develop a multi-lingual, multi-modal search and access system for biomedical information and documents. Several open source components are being integrated into an open architecture for robust and scalable biomedical information search. A number of these components are based on the use of various Language Resources (LRs). These LRs are derived from various heterogeneous knowledge sources. A number of them was proposed at the beginning of the project as listed in Appendix B of the *Description of Work* of the project. This includes text sources such as the Cochrane Systematic Reviews, Open Access Journals, trusted websites and legacy repositories (Clearing House, OMIM...). In addition, image sources are included, such as images from journals, images from openly accessible sources, as well as images from Picture Archiving and Communication Systems (PACS) from two radiology departments in Vienna, Austria and Geneva, Switzerland.

Figure 1 gives a picture of the exploitation cycle of the LRs within and beyond Khresmoi. We can see that LRs used for training and developing the algorithms of the search engine are only used internally, whereas LRs integrated as the knowledge base will be kept in the final search engine. However, knowledge base LRs are encrypted so that they cannot be reconstructed by users of the engine.

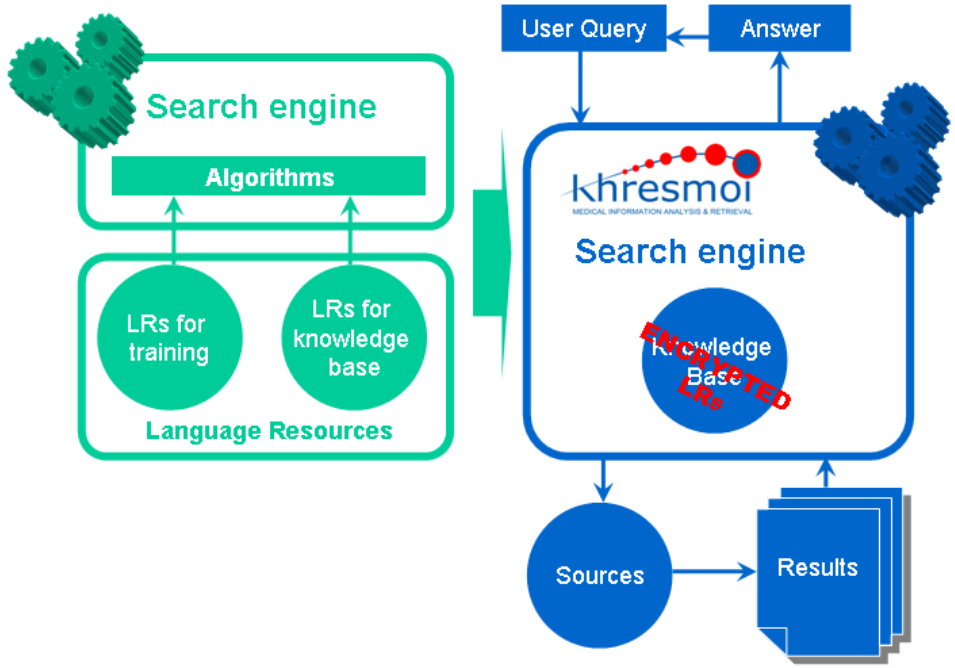

Figure 1: LRs Exploitation cycle

During the project, Language Resources have been used for training purposes but also for evaluating the different components of the search system. The project partners used existing datasets, described in Chapter 2, and produced new datasets presented in Chapter 3 for specific purposes.

In addition to LRs, we identified sources of information to be indexed in the search engine prototype. These are websites containing relevant medical information in the five languages of the project: Czech, English, French, German and Spanish. The identification process is described in Chapter 4.

The last chapter reports on investigation on exploitation rights and licensing issues. We first present an overview of the legal properties that have to be taken into consideration for the exploitation of LRs. This is a presentation of the Intellectual Property in a general context, and the legal framework linked specifically to LRs (public domain, copyright, copyleft, EU database directive, privacy and ethics). We also have to take into consideration the project framework in terms of use limitations (internal versus external use, types of users). Finally, we focus on licensing issues and corresponding rights that currently apply to the LRs identified in the previous sections.

2 Existing Language Resources used within the project

Appendix B in the *Description of Work* of the project gives an extensive list of resources that could be accessed for scientific work within Khresmoi. Throughout the project duration, we went ahead with the identification of other LRs. We also went deeper in the categorization of these LRs with respect to their types and expected uses that are targeted in the project. The list of LRs has been shared with the project partners and was updated regularly through the Khresmoi Wiki pages (http://wiki.khresmoi.eu).

We distinguish three groups of LRs depending on the way they are exploited in the project: one group of LRs used for training systems, a second group used for evaluating components of the Khresmoi system, and the third one used as a knowledge base.

In this context, LRs need to be accessible as a first step to consortium members. In addition, third parties may also be interested in these LRs for their own R&D activities. Rights of use for each of the identified LRs were presented in the Khresmoi wiki.

2.1 LRs for training systems

This group of LRs aims at training and improving the performances of the Khresmoi search engine. It is used within the following tasks:

- Indexing algorithms
- Automated annotation systems
- Classification algorithms
- Extracting translation resources
- Natural language processing algorithms

When possible, the project partners used the identified existing datasets for training their systems. Existing LRs used within the project are categorized below by type of data:

- **Text-centric data sets**, detailed in **Appendix A**: as defined in Khresmoi, these are collections specifically derived from the medical domain. Text-centric data have been used for training the automatic annotation pipeline (WP1). A set of documents from the Genetics Home Reference

data set[1] was manually annotated for medical entities in order to train the annotation team and to better define the methodology [19]. Then web documents have been automatically annotated through the GATE platform[2] and manually corrected for improving the automatic annotation platform. Monolingual and multilingual corpora have also been used for training the Machine Translation (MT) system and to create Language Models (WP4).

- **Image data sets**, detailed in **Appendix B**: although the main aim is to have radiology/radiography images, other types of images have been considered, such as the ImageCLEFmed data, which includes images from articles in the scientific literature. Some images semantically annotated within WP1 have also been used for training the 2D/3D image search system within WP2.
- **Terminologies and ontologies**, detailed in **Appendix C**: these are lists of terms adapted to the medical domain, mainly aimed to be used for the knowledge base work (WP5), but also for annotating the indexed content (WP1), annotating images (WP2) and as domain-specific dictionaries for training MT systems (WP4).

2.2 LRs for evaluating systems

Each component of the search engine has been evaluated in the framework of the evaluation strategy proposed in the deliverable D7.1.1 [9] and D7.1.2 [9]. The evaluation conducted during the second and third year of the project are described and analysed in D7.2 [6]. According to the analysis in [6], 2/5 of the evaluations reused existing resources. Among them, 2/3 came from past campaigns (mainly ImageCLEFmed), while the other 1/3 was related to existing resources not specifically used for evaluation purposes. In the case of reuse of resources, it often happened that formatting or annotations needed to be done to suit the Khresmoi evaluation requirements. The existing LRs used for evaluation within the project are described below:

- **Text-centric data sets**, detailed in **Appendix A**, include :
  - web pages manually classified used for evaluating document categorization through the annotation pipeline (WP1),
  - query logs for evaluating the user categorization system (WP1, T1.5),
  - medical corpora for evaluating the Health relevancy score calculation (WP1, T1.4),
  - monolingual and parallel corpora used for testing the MT system (WP4).
- **Image data sets**, detailed in **Appendix B**, include :
  - test sets of 2D radiographic images (ImageCLEFmed 2009) and images from scientific articles in the biomedical domain (ImageCLEFmed 2011 and 2012) for evaluating the image automatic classification system (WP2),
  - knowing that rights of using medical images are often restrictive, a selected set of images was chosen to be integrated into the Image Search prototype for the users test (WP10).

2.3 LRs as Knowledge Base

The Khresmoi Knowledge Base (KB) is a data warehouse containing a set of semantically integrated data sources specifically tailored for the needs of the project. The data contains both logical

[1] http://ghr.nlm.nih.gov/

[2] http://www.gate.ac.uk

relationships between entities and language resources such as labels and descriptions. Some of the data sources are available in different languages in order to support multilingual queries or for training machine translation algorithms. In addition, implicit relationships between entities are generated based on inference rules, creating additional structured knowledge and providing improved query performance. Details about the modelling of the data are given in [13] and [14].

The KB provides the data and logical relationship for a number of services, which integrate in the project-wide usage scenarios. The disambiguator service is populated with labels, synonyms and descriptions from the KB; the Semantic Type-Ahead Service and Modality Classification service both rely on the data structure in the KB and the inferred logical statements in order to provide the best results; gazetteers for IE are dynamically populated with labels from the KB, which allows them to consistently link terms from unstructured text to entities and keeps them up-to-date. All the services linked to the KB are described in the deliverables [14] and [15].

In the Table 1 we provide a list of the integrated sources and their licensing:

Data source	Language	User profile	License
UMLS MeSH	English, German, French, Spanish, Czech	Radiologists, General public, Medical practitioners	License Agreement for the Use of UMLS Metathesaurus
Radlex	English, German	Radiologists	RadLex® Ontology License
UMLS FMA	English	Radiologists	License Agreement for the Use of UMLS Metathesaurus
UMLS SNOMED	English	Radiologists, General public, Medical practitioners	License Agreement for the Use of UMLS Metathesaurus, specific restrictions (see Section 5.2.5)
Geonames	English	General public, Medical practitioners	Creative Commons Attribution license
DrugBank	English	General public, Medical practitioners	freely available, but redistribution for commercial purposes requires permission [20]
DBPedia	English	General Public	Creative Commons Attribution-ShareAlike 3.0 license and the GNU Free Documentation License

Table 1: Data sources integrated in the Knowledge Base

The size of the Knowledge Base is indicated below (statistics from Oct. 2013):

Number of statements: 1,220,275,956

Number of expl. statements: 917,468,170

Number of entities: 197,390,716

3 Language Resources produced within the project

When no data sets were found, tailored corpora have been created. We consider them as valuable resources, because existing material has been adapted (annotated, anonymized, selected, etc.) in order to create new Language Resources. We categorize them according to the use made during the project: for training or evaluation purposes.

3.1 LRs for training systems

3.1.1 Semantically annotated corpora

A collection of data has been semantically annotated for improving the indexing system. In the context of Khresmoi, the purpose of which is to produce a robust medical search engine, it deals with annotation of documents for medical entities. The aim is to train the automatic classification system for creating the Khresmoi index. Both textual documents and image sets have been manually annotated.

- **The Manually Annotated Reference Corpus**

The Manually Annotated Reference Corpus is a collection of Khresmoi English web documents annotated with key entities (such as disease, drug). It has been constructed by first annotating these entities with an imperfect automatic process through GATE pipelines (described in the Deliverable 1.7 [18]). Then, automatic annotations have been corrected by several human operators. The differences between these human operators were resolved to give the final reference corpus. The corpus is divided into two parts:

1. The initial corpus: 625 documents from the Genetics Home Reference data set [5], automatically annotated with anatomical locations and diseases, and manually corrected by 3-4 annotators. Size of documents: between 26 and 8,306 tokens each.
2. The main corpus: 6,950 English documents from the Khresmoi crawl (see Chapter 4) and 5,518 English Wikipedia pages, automatically annotated through the GATE Platform for Anatomy, Disease, Drug and Investigation. Size of documents: between 200 and 2,000 tokens each. Of these, 1,783 documents have been manually corrected by up to 7 people, to give 68,799 agreed consensus annotations. This final version of this portion of the corpus may be slightly higher, as the final output of the manual annotators is still being integrated.

The production of the Manually Annotated Reference Corpus has been managed by the University of Sheffield (USFD) and manual annotations produced by a team of professional annotators at a Khresmoi sub-contractor, Lighthouse. The corpus is using the GATE XML format. It is fully described in the Deliverable 1.4.1 [17] and final figures will be reported in an updated version of Deliverable D1.4.1.

- **Radiology captions annotated set**

Radiology captions have been manually annotated during the first stage of annotation, which resulted in defining the manual annotation guidelines and management protocol of the large scale, distributed manual annotation effort described in the Deliverable D1.1 [19]. The radiology captions annotated set contains 500 radiology captions from the ImageCLEF2010 Medical Task, annotated by 2 annotators for anatomical parts and diseases. Format: GATE XML format.

3.1.2 Annotated anonymized radiology images

To support the development of the clinical radiology prototype we retrospectively collected a substantial amount of anonymized clinical imaging data (including magnetic resonance imaging, and computed tomography data) from a hospital picture archiving and communication system (PACS). The data was transferred from the *productive system* to the KHRESMOI servers, anonymized, and ingested in a database that allows for continuous experiments, development on parts of the data, and evaluation.

The data comprises approximately 61,000 series, forming more than 4000 studies. For a majority of cases, anonymized radiology reports were also collected, in order to extract volume level labels regarding anatomical structures, and radiological observations made in the data.

The data includes several subsets that serve different purposes in the project:

- **DD1**: a large set of data, for which the inclusion criterium was solely the time range during which the data was acquired. This resulted in a set of consecutive MR and CT data that best reflect the distribution of acquisition techniques, and observations in clinical practice. DD1 was used as the main evaluation and experimentation data for the clinical radiology retrieval algorithms and methods. For a substantial number of volumes, we annotated the location of the centre point in a reference anatomy space. This served as training set for global localization.
- **DD1-addon**: throughout the project DD1 was extended.
- **Radiology reports:** for the majority of cases in DD1 and DD1-addon, we collected and anonymized radiology reports. This data serves as semantic information corresponding to the imaging data. It is used in the form of *weak labels* that describe the image content on a volume level, and as basis for combined image and text retrieval.
- **L1**: lung data with a small set of radiological observations (e.g., emphysema) were collected in order to train and evaluate retrieval methods quantitatively.
- **HRCT:** a set of high-resolution imaging data was collected to serve as basis for initial design and evaluation of local features and image descriptors.

3.1.3 Multilingual corpora from the medical domain

Multilingual data sets have been produced for training the Machine Translation (MT) system:

- **The Khresmoi Query Translation Test Data for the Medical Domain version 1.0**

  This package contains data sets for development and testing of machine translation of medical queries between Czech, English, French, and German. The queries come from general public and medical experts. Size: 1508 medical queries in Czech, English, French, and German.

  The original queries in English were randomly selected from real user query logs provided by Health on the Net foundation (750 queries by general public) and from the Trip database [10] query log (758 queries by medical professionals). Queries have been translated into Czech, German, and French by medical experts.

  The Khresmoi Query Test Set produced within the Khresmoi project is made available under the Creative Commons Attribution-Non-commercial (CC-BY-NC) license, version 3.0 Unported.

- **A Czech/English corpus** of 1,212 manually translated sentences, used during the year 1 of the project to tune the weights in the log-linear system. It contains sentences randomly picked

from pre-chosen sentences (between 20 and 40 words per sentence) from a monolingual English data set. CUNI managed the translation work.

3.1.4 Monolingual corpora from the medical domain

In addition to parallel corpora, domain-specific monolingual datasets have been collected for training the MT system:

- **Czech Biomedical Document Collection**: a collection of 11,410 MeSH-annotated files selected from 25,408 URLs available in the Bibliographia Medica Čechoslovaca (BMČ), version 2011-01. It has been used during the year 1 of the project for training the MT system and during year 3 for evaluating the MT system.
- **Czech monolingual corpus**: a collection of crawled data from Czech websites focused on health and medicine that has been cleaned and POS-tagged. Size: 12,870,939 sentences. It was used to train the Czech Language Model.

3.2 LRs for evaluating systems

Each component of the search engine has been evaluated in the framework of the evaluation strategy proposed in the deliverable D7.1.1 [9] and D7.1.2 [9]. The evaluation conducted during the second and third year of the project are described and analysed in D7.2 [6]. According to the analysis in [6], 3/5 of the datasets used for evaluation have been created within Khresmoi for specific needs. The LRs produced within the project for evaluation purposes are described below:

3.2.1 Semantically annotated corpora

- **Set of reference annotated documents**

A set of 500 randomly chosen web documents crawled by the Health on the Net Foundation (HON) were automatically annotated through the first prototype of the Information Extraction pipeline and manually corrected by the team cited in Section 3.1.1. The original dataset was manually annotated by the HON team and served as a gold standard for evaluating the annotators during their training period. This first annotation stage allowed to improve the performance of the automatic annotation pipeline [1].

- **MeSH annotated corpus in French**

In order to evaluate the first multilingual information extraction prototype, partners of the project were looking for multilingual medical data with gold standard annotations. The ELDA team worked on creating such a dataset in French. A corpus has been collected using a simple mapping between English-French bilingual MeSH terms[1] onto a corpus of PubMed[2] articles written in French (but annotated with English MeSH terms). A total of 4,212 French abstracts have been collected. The collection of documents covers 23,014 unique MeSH terms among 72,929 assignments, which represent more than 17 MeSH terms per document. For copyright reasons, the corpus contains abstracts only. Negotiations would be needed before providing the full text documents associated to the abstracts.

[1] http://mesh.inserm.fr/mesh

[2] http://www.ncbi.nlm.nih.gov/pubmed

3.2.2 Annotated anonymized radiology images

The radiology images from the hospital picture archiving and communication system (PACS) described in Section 3.1.2 were also used as evaluation material. Evaluation was performed in a cross validation fashion by training the system on a subset of the data, and applying the algorithms to the remaining cases. This procedure was iterated, so every part of the data served as training and test set at least once. The data comprises the following subsets:

- **DD1** (large set of consecutive MR and CT data), used as the main evaluation and experimentation data for the clinical radiology retrieval algorithms and methods.
- **DD1-addon** (extension of DD1).
- **L1** (lung data with radiological observations), used to evaluate retrieval methods quantitatively.
- **HRCT** (set of high-resolution imaging data), used to evaluate local features and image descriptors.

3.2.3 PubMed Central images dataset

A dataset has been produced using the PubMed Central Open Access Subset[1] by extracting the subfigures out of the compound figures and automatically annotating with regard to their image modality. This resulted in a dataset of 1.7 million images. In this dataset information also exists about the image captions, the image parent (if they are subfigures) and the corresponding article the image exists in. This dataset is planned to be used for the last round of user-oriented evaluation (WP10).

3.2.4 Benchmark dataset for Information Retrieval in the medical domain

The **CLEF eHealth 2013 Task 3 Evaluation Package** contains the information retrieval test collection produced for the CLEF eHealth Lab 2013, Task 3[2], held in collocation with CLEF2013, Valencia, Spain. Task 3 of the Lab aimed at facilitating the evaluation of information retrieval techniques for the medical domain, specifically information retrieval to address questions patients may have when reading clinical reports.

The data collection consists of a set of one million medical-related documents, provided by the Khresmoi project, covering a broad set of medical topics. The documents in the collection come from several online sources, including Health On the Net organization certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia).

Sample development queries and result sets are provided with the data set. The queries have been manually generated by medical professionals from a manually extracted set of highlighted disorders identified in Task 1 of CLEF eHealth 2013. Relevance assessment has been performed by medical professionals. The creation of the data sets is fully documented in [5].

This collection was created in the framework of Khresmoi for evaluating search by the general public. It is planned to be used for evaluating textual search & ranking during the last year of the project (see D7.1.2 [9]). All the data sets and guidelines produced for the CLEF eHealth Task 3 Evaluation Challenge, along with results from participants are provided in the Evaluation Package available

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[2] http://www.nicta.com.au/business/health/events/clefehealth_2013

through the ELRA catalogue (http://catalog.elra.info): CLEF eHealth 2013 Task 3 Evaluation Package, catalogue reference: **ELRA-E0042**.

3.2.5 Multilingual corpora from the medical domain

Multilingual data sets have been used for evaluating the MT system:

- **The Khresmoi Query Translation Test Data for the Medical Domain version 1.0**, described in Section 3.1.3, was also used for evaluating the MT system.
- **The Khresmoi Summary Translation Test Data for the Medical Domain**, currently being created for evaluating the summary translation system (see D7.1.2 [9]). It will contain sentences from summaries of medical articles in Czech, English, French, and German. The original sentences are sampled from automatically generated summaries of English medical documents crawled from the web in 2012 (selected from the CLEF 2013 eHealth Task 3 collection) and identified to be relevant to 50 medical topics provided for the same task. Sentences were extracted from a set of websites allowing use of their content for producing derivative work. The Khresmoi Summary Test Set will be made available under the terms of the Creative Commons Attribution-Non-commercial (CC-BY-NC) license, version 3.0 Unported.

The two datasets described above will also be provided as development and test data for the WMT 2014 medical translation task[1].

- **A Czech/English corpus** of 956 manually translated sentences, used in the first year of the project for evaluating the MT system. It contains sentences randomly picked from pre-chosen sentences (between 20 and 40 words per sentence) from monolingual English data. CUNI managed the translation work.

3.2.6 Monolingual corpora from the medical domain

The **Czech Biomedical Document Collection**, described in the Section 3.1.4, was also used for evaluating the MT system during the third year of the project.

3.2.7 Sets of queries from logs

We will describe here sets of query logs manually selected and adapted for evaluating components of the search engine. Other sets of raw query logs used for evaluation are described in Appendix A.

- General Public query sets:
    - **Set of 50 English 'short' queries from General Public logs**: the 50 general public queries were manually selected from a sample of raw queries of HON search engines collected over a period of 6 months (Nov. 2011 – May 2012). Only non-capitalized queries were taken into account to remove possible influence by web crawlers using predetermined queries. The 50 English queries were selected by a domain expert (Natalia Pletneva, HON). Queries which seemed to be too "medical" and in languages other than English were excluded [1]. This data set has been used during the first

[1] http://www.statmt.org/wmt14/medical-task/

years of the project for the query set analysis described in D1.3 [1], and for the pilot global empirical evaluations conducted in year 3 (see D7.2 [7]).

- **Set of 50 English 'long' queries from General Public logs**: an additional set of queries which contains more than 2 terms were selected from a sample of HON raw queries collected from Nov 2011 to Jan 2013. This data set has been used for the pilot global empirical evaluations conducted in year 3 (see D7.2 [7]).
- **Set of 100 multilingual queries from General Public logs**: this is a collection which contains the English queries described above (short and long queries), along with their manual translation into Czech, French and German. This dataset was built to evaluate the query translation service for the pilot global empirical evaluations conducted in year 3 (see D7.2 [7]).
- **Set of 100 multilingual queries with spelling errors from General Public logs**: in order to evaluate the spelling corrector service, spelling errors were added to the dataset of multilingual queries described above. Spelling errors include diacritics omission, leaving out white space, character omission, character insertion, character replacement, and character swapping. This dataset was built for the pilot global empirical evaluations conducted in year 3 (see D7.2 [7]).

- **Set of queries from General Practitioner logs**: the 50 English general practitioner queries were manually selected by domain experts (Matthias Samwald from TUW, Marlene Kritz from GAW) to represent a variety of common search phrases found in the available query logs (PubMed query log, Trip database query log). Queries that contained spelling mistakes or which seemed not to stem from clinical information needs of medical professionals were excluded. This dataset was used for query set analysis (see D1.3 [1]).
- **Set of 5 queries** extracted from the 50 'long' General public logs: for each of the 5 queries, a set of 5 relevant and 5 non-relevant documents was manually generated with the Khresmoi prototype system. This set of queries along with relevant and non-relevant documents has been used to evaluate the summarization system (see D4.4 [11]).

4 Data sources indexed in the current prototype

In order to test the Khresmoi search engine prototypes, websites considered as essential when looking for medical information have been identified in the different languages of the project: Czech, English, French, German and Spanish. The Khresmoi search system developed during the first years of the project [8] being based on an existing search system at the Health on the Net Foundation (HON), the aim was to complete the existing index of HON-certified websites. Websites indexed by HON are regularly certified according to their compliance to the HON Code of Conduct for medical and health Web sites (HONcode)[1].

A list of non-certified websites to be added into the Khresmoi index was settled and updated by the project partners. The list includes data sources such as the European Medicines Agency[2], the Ministry of Health of the Czech Republic[3], and other major sources of medical information in the countries involved in the Khresmoi project.

[1] More information on the HONCode principles can be found at http://www.healthonnet.org/HONcode/

[2] The website of the European Medicines Agency: http://www.ema.europa.eu

[3] The website of the Czech Ministry of Health: http://www.mzcr.cz

Then specific partners involved in users tests (WP10) have been assigned to refine the list according to the feedback from different users targeted by the project [2] (general public, medical professionals):

- The Society of Physicians in Vienna (GAW) was in charge of web sources for professionals,
- Health on the Net (HON), in charge of web sources for patients,
- Charles University in Prague (CUNI), for both types of users in Czech.

For each Web source, partners of the project had to indicate:

- Name of the Resource
- URL start page (that is where the crawl will start)
- Most Critical Resource [X or nothing]
- Classification 1 [for_patients/ for_professional/for_both]
- Classification 2 [Clinical Practice / Definition / Medical Education / Medical Forum / News / Organisational / Patient / Scientific Search]
- Sub Classification [Medical education: Online cme / Event, Clinical Practice: Guidelines / Drug information / Diagnosis]
- Language
- Comments in free text

The file completed during this first stage of identification is shared via a Google document, which allows the Khresmoi Crawling system to automatically access the content for crawling and categorizing the data. The final workflow for crawling data sources is described in the Deliverable 1.7 [16], along with the processing, semantic indexing and ranking of documents.

At the time of writing this Deliverable (February 2014), a final number of 2,732,430 web pages from 193 sources have been crawled by HES-SO, the project partner in charge of crawling the websites not certified by HON. These web pages will be added to the Khresmoi index, along with web pages from the 8,300 HON-certified websites.

A list of the 40 most important websites crawled from the list identified by project partners is provided in Appendix D.

It must be noticed that availability of data is not constant. Although identified in Appendix B of the *Description of Work* of the project, some data sources disappeared during the duration of the project. For instance, the medical encyclopedia Medpedia (www.medpedia.com) no longer exists at the time of writing[1].

We could also consider the list of websites manually classified described above as a useful input for evaluating, for instance, an automatic classification system or for creating medical content indexes in the Information Retrieval domain.

5 Investigation on exploitation rights

For the Khresmoi project consortium, it is crucial that the permission to use copyrighted LRs is obtained. As an introduction to this question, we will first present some key aspects of legal issues. Then, we focus on licensing issues and corresponding rights that are applicable to the LRs defined in the previous section.

[1] A report on the death of Medpedia is accessible at http://www.crunchbase.com/company/medpedia

5.1 Legal issues

5.1.1 Copyright, Intellectual property, and EU Database Directive

The Language Resources used within the activities carried out by the Khresmoi Consortium refers to electronic data such as textual corpora, terminology resources, images, etc. Such datasets are produced by third parties that are hence the right holders. Such ownership is often governed by copyright laws (very rarely by patent laws). In the Khresmoi context, resources that are copyrighted by third parties are usable within Khresmoi either because they have been negotiated for the use within the project and for further distribution or because the owners decided to licence them under more permissive licences covering the Khresmoi usages. In several cases, the copyrighted material is licensed under some implicit and permissive licensing schema such as Creative Commons, that allows such used without negotiations.

The Directive 96/9/EC of the European Parliament and the Council of 11 March 1996 has defined a specific sui generis right for databases: this right is irrespective of the innovative intrinsic nature of this work, but made to help investors safeguard their investments on collecting and compiling such data. This directive assesses in particular that the re-utilization or extraction of the contents of a database can have a serious economic and technical impact.

5.1.2 Privacy and Ethics

The last crucial aspect that has been considered when manipulating our resources is related to Privacy and Ethical issues. Personal data are sometimes necessary to produce Language Resources. In some cases, personal data can even be part of the LRs (e.g. medical reports with names and addresses, date of birth, etc.). This issue is even more sensitive when collecting, storing, using and exchanging biometric data, due to the difficulty of keeping the anonymity of persons.

At the European level, this issue was taken into consideration in particular in the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on “the protection of individuals with regard to the processing of personal data and on the free movement of such data”.

Among the requirements related to the use of personal data, we can quote in particular:

- a fair and lawful processing of personal data,
- well specified, explicit and legitimate purposes,
- accuracy of collected information and adequacy to the purposes,
- storage of information on a reasonable/not excessive period of time.

In addition to this Directive, Article 8 of the Charter of Fundamental Rights of the European Union (18/12/2000), entitled “Protection of personal data and states”, assesses the following:

1. Everyone has the right to the protection of personal data concerning him or her.
2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
3. Compliance with these rules shall be subject to control by an independent authority.

Furthermore, Article 10 of Regulation (EC) 45/2001 states that “the processing of special categories of data, defined as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and of data concerning health or sex life, is prohibited, subject to certain exceptions”.

As far as Khresmoi is concerned, LRs that are compiled for the project may contain information related to personal data or information touching people privacy. In both cases, pre-processing has **ensured that personal data is anonymized while "private" parts are deleted.**

5.1.3 Legal framework for LRs: Public domain, Copyright and Copyleft

From the investigation carried out within the Khresmoi project, we identified three main scenarios that should be distinguished for the use of LRs.

The first scenario is related to "Public Domain data". This is in fact an "IPR-free" category (i.e. any ownership is left out). Works are in the public domain and they are not covered by intellectual property rights at all (the intellectual property rights have expired and/or the intellectual property rights are forfeited). This is the ideal category for Khresmoi, as public domain data can be freely used for derivative works and further distribution without permission.

The second scenario is about "Copyleft". It is unofficially represented by a reversed "c" in a full circle (in opposition to the copyright symbol). It is a method to make works and their re-use freely available as long as the same rights that apply to the original work shall be kept with the derivative works. Several licenses, exploit this copyleft modality, such as the GNU GPL (General Public License) or the Creative Commons Attribution-Share Alike license (CC BY-SA).

The third scenario is related to copyrighted resources. The copyright law grants any creator or author of the original work (physical or moral entity) the rights to copy, distribute and adapt the work. These rights cannot be waived without the right owner's consent or in special circumstances authorized by the copyright law. Examples of special circumstances that do not require the permission of the copyright owner, as given by the WIPO (World International Property Organization), are:

- quoting from a protected work, provided that the source of the quotation and the name of the author is mentioned, and that the extent of the quotation is compatible with fair practice;
- using works by way of illustration for teaching purposes; and
- using works for news reporting.

In the US, exceptions for research and education purposes are also considered under the terms "fair dealing" and "fair use".

When a copyright law applies, it is an absolute necessity that each user seeks prior agreement from the rights holder (or buys such rights from the LR owner).

Beyond these three modalities, we can also mention the existence, in some jurisdictions such as France, Germany, of moral rights that include different rights (from the economic rights) such as the right of attribution, the right to the "integrity" of the work, etc. The moral rights are directly assigned to the creator and are inalienable and not transferable under a number of jurisdictions. However, such rights are unlikely to affect the LR issue within this project and within the Human Language Technology (HLT) field to a larger extent.

5.1.4 Limitations on use

Two main questions should be considered to define the limitations on use of all LRs used by Khresmoi partners: What are the types of use? What are the types of users?

5.1.4.1 Types of use

As a preliminary work, it is important to investigate how LRs are to be exploited within the project. This investigation is part of the identification work. From this investigation, we could highlight two main types of use, "internal use" and "external use", as defined below:

- "Internal use": we can observe that all the tasks defined in section 2.1 are limited to an exploitation of LRs within the project. No LR will be used outside the project. From ELDA's experience, this kind of use can be easily associated with research and development of technologies. This is crucial to have this in mind when discussing with the LR providers.

- "External use": as presented in section 2.3, the idea is to provide a derivative product in the form of the Khresmoi search engine. This derivative product will exploit the original LRs and will be provided in the form of a package that includes the LRs when being distributed to third parties. However, the LRs will necessarily be encrypted and non reversible, so that no end-user may be able to reconstruct the original LR. Moreover, we envisage an external use, which is the distribution of value-added resources. As presented in section 3, value-added resources are LRs that have been used within the project and have been modified to be adapted to the project requirements. Some of those LRs will be packaged and distributed to third parties who would like to test, train and develop their own tools.

5.1.4.2 Types of users with respect to types of use

Having in mind the types of use, we also need to clarify who will be the users for both situations (internal versus external use). Indeed, this information is crucial during the negotiation with rights holders of LRs.

As far as the internal use is concerned, it is quite clear that the users are the project consortium members. The Khresmoi partners' relationship is bound and governed by a consortium agreement. This facilitates the signature task as one single entity can obtain the ability to sign on behalf of the whole consortium to avoid multi-part agreements.

Regarding the external use, it is important to define who will be granted the right to distribute the (value-added) LRs. In a general way, users can be either laboratories involved in public or private research on Human Language Technologies (HLT) or companies involved in technology development or deployment.

5.2 Licensing issues

In order to use the LRs within Khresmoi, the best approach is to be granted a licence for these purposes by the right holders.

To use the identified LRs within Khresmoi, we had to check the feasibility to use them in conformance with the authorized and/or granted rights. Thus, for each LR, we identified the associated rights as reflected in the licensing schema indicated by rights holders. Beyond public domain and copyrighted LRs, we had to identify resources governed by the following licenses: GNU, Creative Commons, ELRA Language Resources licenses and specific licenses. These are described below. An extensive overview on licensing issues can also be found in deliverable 6.1.1 from the T4ME Net (META-NET) project [3].

5.2.1 GNU Free Documentation License (GNU FDL)

The GNU Project[1] was initiated in 1984, in order to offer a free software operating system as an answer to the paying and proprietary Unix operating system (GNU is a recursive acronym standing for "GNU's Not Unix"). The first version of GNU licenses appeared in February 1989[2]. From the original project, several licenses emerged: GNU General Public License, GNU Lesser General Public License, GNU Affero General Public License, GNU Free Documentation License (GNU FDL). From our investigation, we identified several LRs that are protected by the GNU FDL. Therefore we are limiting our analysis to this specific license only. The main specificities of this license are based on the following criteria:

- **Reuse**: this license allows any user of related material to "copy and redistribute it, with or without modifying it, either commercially or non-commercially".
- **Freedom**: everyone using the material referred to in this license can do it in the same sense of freedom as indicated in the license.
- **Copyleft**: any material deriving from the material referred to in this license shall keep the same conditions of use as stated in the original license.
- **Authorship preservation**: the license preserves credits to the original authors/publishers as well as requires anyone that modifies the original LR both to make credits to the original authors/publishers and endorse responsibilities for the related modification.

5.2.2 Creative Commons

Creative Commons[3] is a non-profit organization which was created in 2001 with the spirit of sharing creative works and offering another option to "all rights reserved" copyright. To this aim, it worked out several licenses, named Creative Commons (CC) licenses.

As far as the LRs identified for Khresmoi are concerned, we can highlight the following ones:

- **Creative Commons Attribution license (CC BY)**: this license allows others to freely copy, distribute, remix, tweak, and build upon one's work, even commercially, under the "Attribution" condition. In this sense, "attribution" means that anyone using a work provided under this license shall "attribute the work in the manner specified by the author or licensor". No endorsement is implied though. The main requirement for any reuse or distribution is that anyone using LRs under this license shall make clear to others the license terms of the work. This can be done by linking to the license page (http://creativecommons.org/licenses/by/3.0/).

- **Creative Commons Attribution-Share Alike License (CC BY-SA)**: this license has the same basis of rights and constraints as the CC BY license. However, it includes a "share alike" condition: any new creation shall be licensed under the identical terms as the original license. This license is often compared to "copyleft" free and open source software licenses. Any derivative work will then carry the same rights as the original works. Again, the main requirement for any reuse or distribution is that anyone using LRs under this license shall make clear to others the license terms of the work. This can be done by linking to the license page (http://creativecommons.org/licenses/by-sa/3.0).

[1] http://www.gnu.org

[2] http://www.gnu.org/licenses/old-licenses/gpl-1.0.html

[3] http://creativecommons.org

- Another variant that could be of importance for Khresmoi is the **Creative Commons Attribution-Non Commercial License (CC BY-NC)**: this license means that no use of the "work" shall be done within commercial products. This is too restrictive since Khresmoi targets market exploitation of the technology it develops, including the search engine.

5.2.3 ELRA license

ELRA[1] (European Language Resources Association) was created in 1995 to make available Language Resources for Human Language Technologies and to evaluate such technologies. Among ELRA's works, one of the original ones was to draft generic agreements to enable the sharing of Language Resources. ELRA's licenses allow both research and commercial activities through a number of constraints that in particular prevent LRs from being redistributed as such (it should be encrypted in a way that the original LRs shall not be rebuilt).

ELDA, acting as a distribution agency for ELRA, gives the possibility to LR providers to distribute their resources on the basis of two features: user profile (academic vs. commercial) and purpose of use. The licenses between ELRA and the users grant the latter a non-exclusive, non transferable right to use, rework and build on the Language Resources for the purposes agreed upon between the provider and ELRA within the user's institution. In this classification, three types of licenses are foreseen:

- **Language Resources End-User Agreement:** this license allows research activities conducted by any organization including for-profit ones. Licensees are granted "the right to rework and build upon the Language Resources, or any component thereof, as necessary or desirable for the purposes of their own internal language engineering research activities." However, no distribution or marketing of any derivative product or service based on the original Language Resources is authorized.

- **Language Resources VAR Agreement**: Value-Added Resellers (VARs) are users who are authorized to create derivative works or software for internal research (such as the End-User agreement), as well as for the development of products and services. They may distribute derivative products resulting from their development activities, whereas they are not "permitted to make available to the public all or any substantial part of the contents of the Language Resources".

- **Evaluation Packages End-User Agreement**: a number of resources have been designed for evaluation and benchmarking to assess the performance of Human Language Tools and Technologies. ELRA is distributing them in the form of "Evaluation Packages" that shall be used solely for this purpose. Other uses are not permitted.

5.2.4 Radlex License

Among LRs that were identified, one of them is provided through a proprietary license: the RadLex® Ontology. RadLex® is a reference ontology for the domain of radiology, produced within the RadLex® still ongoing project through a funding of the Radiological Society of North America (RSNA).

[1] http://www.elra.info

Each release version of RadLex® is free of charge and can be obtained by registering through a website and by accepting and signing the corresponding license[1]. The license reads in particular: "This License permits public access to the Release Version of RadLex® and makes it possible for Licensees to use it for clinical, research, educational and commercial activities without charge".

The license "grants Licensee a perpetual, worldwide, non-exclusive, no-charge, royalty-free, copyright license to reproduce, publicly display, publicly perform, prepare Modifications, and distribute the Work with or without Modifications". Any licensee shall follow the conditions mentioned in article 4 of the license. In particular, references to the RadLex® ontology shall be respected, and the licensee shall provide all recipients of the Work or Modifications a copy of the License.

5.2.5 UMLS Metathesaurus license

According to the Metathesaurus License[2], the vocabularies from the UMLS Metathesaurus is *freely available and can be incorporated in any computer applications or systems designed to improve access to biomedical information of any type*, subject to the main following restrictions:

1. full **redistribution** of vocabulary sources is **forbidden** (see article 3 of the full license),
2. it is required to **inform the National Library of Medicine** (NLM) prior to **distributing** any application(s) in which it is using the UMLS Metathesaurus (see article 4),
3. provide NLM with a **brief report** on the usefulness of the UMLS Metathesaurus **each year** (see article 5),
4. **acknowledge NLM** as its source of the UMLS Metathesaurus, **citing the year and version number** (see article 10),
5. if using any material from the UMLS Metathesaurus which is from other copyrighted sources, display the list of vocabularies with any appropriate copyright notice and whether the entire contents are present or only a portion of it. The following sentence should be stated along with the vocabularies: "*Some material in the UMLS Metathesaurus is from copyrighted sources of the respective copyright holders. Users of the UMLS Metathesaurus are solely responsible for compliance with any copyright, patent or trademark restrictions and are referred to the copyright, patent or trademark notices appearing in the original sources, all of which are hereby incorporated by reference*."
6. when using material from the UMLS Metathesaurus which is from other copyrighted sources, **additional restrictions** may apply. Such restrictions depend on the "Restriction Category" of each vocabulary. Some of the vocabularies restrict the use of data for internal use at the licensee location, some others forbid the release of any derivative work using the vocabularies.

There is also a specific license for SnomedCT, which restricts the use of data depending on the territory where each organization is located (people in non-member countries must pay a fee). In order to ask for the right of using data as a Consortium, a specific License for Research project was requested.

5.3 Negotiations

A negotiation process has been launched with LR owners/providers. Negotiations were conducted by ELDA and other partners from the Khresmoi project.

[1] http://www.rsna.org/radlex

[2] https://uts.nlm.nih.gov/license.html|UMLS

It pertains to datasets that needed an (explicit) signed license or datasets for which a special license for research purpose had to be negotiated.

Thanks to identification of right of use for each selected dataset, a large part of the LRs used had a permissive licensing schemas and did not need any negotiations.

Agreements have been obtained (for internal use only) for the LRs described below:

- Cochrane collection of systematic reviews, provided by the Cochrane Collaboration[1],
- The French MeSH 2012, provided by the French Institute of Health and Medical Research (INSERM),
- A set of search logs from the TRIP database (http://www.tripdatabase.com/),
- A radiology book available in electronic format (also called Radio2wiki): "Lehrbuch der radiologisch-klinischen Diagnostik", Prof. Dr. G. LECHNER und Prof. Dr. M. BREITENSEHER, 2003.

Negotiations are still ongoing for a few other data sources.

When crawling the data for the Khresmoi index (see Chapter 4), a few websites did not allow their content to be crawled. After contacting them, a majority of them granted the Khresmoi project the right to crawl their data, knowing that the Khresmoi search engine prototype will redirect the end-user to the source website. No further use will be made of their data. The websites who authorized Khresmoi to crawl their content are listed in Table 2.

URL	Source
www.dh.gov.uk/en	Department of health, NHS news, local UK ;
www.signaling-gateway.org	UCSD signalling pathway
www.eaccme.eu	European Accreditation Council for Medical Education
http://www.medscapemedizin.de	Medscape Deutschland
http://praxis.medscapemedizin.de	Medscape Medizinpraxis Deutschland
www.attract.wales.nhs.uk	ATTRACT

Table 2 : Websites who granted Khresmoi the right to be crawled

Negotiations are still ongoing for several other websites.

5.4 Sharing LRs after the project

Some of the value-added resources created within the project have already been released:

- The CLEF eHealth 2013 Task 3 Evaluation Package, released under the ELRA Evaluation Packages license, available from the ELRA catalogue[1] under the reference: **ELRA-E0042**.

[1] http://www.cochrane.org/

- The Khresmoi Query Translation Test Data for the Medical Domain version 1.0, released under Creative Commons Attribution-Non-commercial (CC-BY-NC) license, available from the LINDAT/CLARIN repository[2].
- The Khresmoi Summary Translation Test Data for the Medical Domain, released under Creative Commons Attribution-Non-commercial (CC-BY-NC) license, will be made available soon via the LINDAT/CLARIN repository.

Discussions are being held to make more LRs available beyond the project.

6 Conclusion

As a conclusion we can say that a large range of existing data sets has been used within the Khresmoi project, from text-centric datasets to images and terminological resources. The focus was put on medical related content. Although this is a growing research field, we noticed a fair amount of data was found in the English language, but it was more difficult to find medical-related datasets in French, German, Spanish and Czech, especially for corpora annotated with medical entities.

In addition, value-added resources were created from existing data. LRs were produced either when no existing data sets were found or for training the annotation teams. A large effort has been made for annotating medical-related documents in categories for improving the automatic classification system. This resulted in a Reference Corpus that could be used for evaluating such systems. Several monolingual and bilingual datasets were produced for training and evaluating the MT system and multilingual information retrieval. A benchmark dataset created for evaluation of General Public queries has been created in the framework of an evaluation challenge in Information Retrieval, the CLEF eHealth Lab (Task 3), held in 2013 and 2014. Finally, some datasets of queries from logs have been extracted and adapted for different evaluation purposes.

Now, the work done by checking legal and licensing issues will save some time when finalizing discussions on LRs that could be shared beyond the project or not.

[1] http://catalog.elra.info

[2] http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5

7 References

[1] Niraj Aswani, Liadh Kelly, Mark Greenwood, Angus Roberts, Matthias Samwald, Natalia Pletneva, Gareth Jones, Lorraine Goeuriot (2012). D1.3 Report on results of the WP1 first evaluation phase. Khresmoi public deliverable.

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

[3] Khalid Choukri, Stelios Piperidis, Prodromos Tsiavos, John Hendrik Weitzmann (2011), Legal, IPR and Licensing Issues in Meta-Share – Meta-Share Licences, Deliverable 6.1.1 of T4ME Net (META-NET) project, http://www.meta-net.eu.

[4] Andreas Eisele and Yu Chen (2010). MultiUN: A Multilingual corpus from United Nation Documents, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC), La Valletta, Malta, European Language Resources Association (ELRA).

[5] Genetics Home Reference: http://ghr.nlm.nih.gov/

[6] Lorraine Goeuriot, L. Kelly, Gareth J. F. Jones, Guido Zuccon, Hanna Suominen, Allan Hanbury and Henning Muller (2013). Creation of a New Medical Information Retrieval Evaluation Benchmark Targeting Patients Needs. In proceedings of The Fifth International Workshop on Evaluating Information Access (EVIA 2013), a Satellite Workshop of the NTCIR-10 Conference.

[7] Lorraine Goeuriot, Olivier Hamon, Allan Hanbury, Gareth Jones, Liadh Kelly, Johannes Leveling, Joshua Robertson, Ales Tamchyna (2013). D7.2 Meta-analysis of the first phase of empirical and user-centered evaluations. Khresmoi public deliverable.

[8] Allan Hanbury, William Belle, Nolan Lawson, Ljiljana Dolamic, Natalia Pletneva, Matthias Samwald, Célia Boyer (2012). D8.3 Prototype of a first search system for intensive tests. Khresmoi public Deliverable.

[9] Liadh Kelly (2012). D7.1.1 User-centered and empirical evaluation strategy and recommendations. Khresmoi internal deliverable.

[10] Liadh Kelly (2013). D7.1.2 Updated user-centered and empirical evaluation strategy and recommendations. Khresmoi internal deliverable.

[11] Liadh Kelly, Johannes Leveling, Shane McQuillan, Sascha Kriewel, Lorraine Goeuriot, Gareth Jones (2013). D4.4 Report on summarization techniques. Khresmoi public Deliverable.

[12] Emma Meats, Jon Brassey, Carl Heneghan, and Paul Glasziou (2007). Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? Journal of the Medical Library Association. 95(2), 156–163.

[13] Konstantin Pentchev, Vassil Momtchev (2011). D5.1 Report on data source integration. Khresmoi public Deliverable.

[14] Konstantin Pentchev, Vassil Momtchev (2012). D5.2 Large Scale Biomedical Knowledge Server. Khresmoi internal Deliverable.

[15] Konstantin Pentchev, Vassil Momtchev (2013). D5.4 Report on consistency checking rules for information extraction. Khresmoi public Deliverable.

[16] Natalia Pletneva, Sacha Kriewel, Marlene Kritz (2013). D8.5.2 Prototype of a second search system based on feedback. Khresmoi public deliverable.

[17] Angus Roberts, Johann Petrak, Niraj Aswani (2013). D1.4.1 Report accompanying Manually Annotated Reference Corpus. Khresmoi public deliverable.

[18] Angus Roberts, Johann Petrak, Célia Boyer, Ljiljana Dolamic, Allan Hanbury, Michael Dittenbach, and Julien Gobeill (2014). D1.7 Prototype and report on semantic indexing and annotation for information retrieval. Khresmoi public deliverable.

[19] USFD Team. D1.1 Manual annotation guidelines and management protocol (2012). Khresmoi internal deliverable.

[20] Wäschle, K. and Riezler, S. (2012b). Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pp. 12-27.

[21] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. (2008) Jan; 36 (Database issue): D901-6. PMID: 18048412

Appendices

A. List of text-centric data sets used for training / evaluation

Data set	Description	Training	Evaluation
AOL logs (or 500k User Session Collection)[1]	Logs obtained from March to May of 2006 and divided into two non overlapping sets: AOL-Health and AOL-NotHealth		WP1, T1.5
ARRS GoldMiner[2]	Logs from an image search engine that provides access to more than 300,000 radiology images based on text queries of text associated with the images.		WP1, T1.5
Bibliographia Medica Čechoslovaca (BMČ)[3]	Collection of references to selected articles for medical specialists (25,408 URLs)	WP4	
Brown Corpus[4]	Corpus of one million words of American English texts printed in 1961		WP1, T1.4 (see [18])
CESART Evaluation package (ELRA-E0019)[5]	Material that was used for the CESART evaluation campaign (evaluation of terminology extraction tools). Includes a medical corpus of web pages (7,514 documents / 9,000,000 words) extracted from Santé Canada.	WP4	
CESTA Evaluation Package (ELRA-E0020)[5]	Material that was used for the CESTA evaluation campaign (evaluation of MT systems). Includes English-French parallel corpora in the medical domain.	WP4	
Cochrane[6]	Large collection of high-quality systematic reviews.	WP1, WP4	

[1] Obtained from http://www.gregsadetsky.com/aol-data/

[2] http://goldminer.arrs.org

[3] http://www.medvik.cz/bmc/index.do?language=en

[4] http://icame.uib.no/brown/bcm.html

[5] http://catalog.elra.info

6 http://www.cochrane.org/

ComCrawl	A corpus of web crawl data composed of over 5 billion web pages. This data set is freely available on Amazon S3 and is released under the Common Crawl Terms of Use. The CZ-EN, DE-EN and FR-EN subsets have been used.	WP4	
COPPA (Corpus Of Parallel Patent Applications)[1]	Contains English-French Patent Cooperation Treaty applications (title and abstract) published between 1990 and 2010	WP4	
ECDC translation memory[2]	Translation memory produced by the European Centre for Disease Prevention and Control. Collection of health-related documents with professional translations into 25 languages.	WP4	
EMEA corpus[3]	Parallel corpus composed of documents from the European Medicines Agency.	WP4	WP4 (a subset)
EQueR Evaluation Package (ELRA-E0022)[4]	Material used for the EQueR evaluation campaign (evaluation of Question-Answering systems). Includes a medical corpus of about 140 Mb of data consisting of scientific articles and guidelines for good medical practice, selected by CISMeF (Catalogue et Index des Sites Médicaux Francophones) from the University Hospital Centre of Rouen.	WP4	
Europarl[5]	Parallel corpus extracted from the proceedings of the European Parliament. Includes EN, FR, DE, and CS documents	WP4	
Genetics Home Reference [5]	Consumer-friendly information about the effects of genetic variations on human health.	WP1, WP4	
GENIA corpus[6]	Collection of biomedical literature compiled and annotated within the GENIA project. Contains 1,999 Medline abstracts, selected using a PubMed query for the three MeSH terms "human", "blood cells", and "transcription factors"	WP1, WP4	

[1] http://www.wipo.int/patentscope/en/data

[2] http://ipsc.jrc.ec.europa.eu/?id=782

[3] http://opus.lingfil.uu.se/EMEA.php

[4] http://catalog.elra.info

[5] http://www.statmt.org/europarl/

[6] http://www.nactem.ac.uk/genia/

GigaWord	A subset from the LDC2003T05 Gigaword corpus[1], a corpus of newswire text data in English	WP4	
GREC[2] (Gene Regulation Event Corpus)	The GREC corpus is a semantically annotated corpus of 240 MEDLINE abstracts (167 on the subject of E. coli species and 73 on the subject of the Human species) which is intended for training IE systems and/or resources which are used to extract events from biomedical literature.	WP4	
Hansard French/English (LDC95T20)[1]	Drawn from official records of the proceedings of the Canadian Parliament	WP4	
HON-certified websites	A subset of English, French, German and Czech web documents among the HON certified websites.	WP4	WP1 (subset of 500 documents), WP7
HON classified diabetes websites	A subset of web documents classified in the 'diabetes' section	WP1, WP4	
JRC-Acquis[3]	Parallel corpus extracted from Acquis Communautaire, the total body of European Union law applicable in its Member States, by the Language Technology group of the European Commission's Joint Research Centre. Includes EN, FR, DE, and CS parallel documents.	WP4	
Linguee	52 German/English aligned sentences from the online dictionary service Linguee[4].	WP4	
MEDLINE abstracts	~20 millions abstracts covering all literature in health and life sciences	WP1, WP4	
MuchMore Springer Bilingual corpus[5]	Parallel corpus of English-German scientific medical abstracts obtained from the Springer web site. The corpus was aligned on the sentence level. A tagged version of the corpus	WP4	WP1, T1.4 / WP8

[1] http://catalog.ldc.upenn.edu/

[2] http://www.nactem.ac.uk/GREC/

[3] http://www.jrc.it/langtech

[4] http://www.linguee.de/

[5] http://muchmore.dfki.de/resources1.htm

	is available.		
MultiUN [4]	A multilingual parallel corpus extracted from the official documents of the United Nations. The English and French parts have been used	WP4	
News Commentary parallel corpus[1]	WMT 11 version available for the following language pairs: FR–EN, DE–EN, and CS–EN	WP4	
OJEU	Official Journal of the European Union in Formex 4 (XML) from the middle of 2004 to the beginning of December 2010 in up to 23 languages. Within Khresmoi, the CZ-EN, DE-EN and FR-EN subsets have been used.	WP4	
PatTR	A sentence-parallel corpus extracted from the MAREC patent collection. The German-English subset has been used.	WP4	
PIL (Patient Information Leaflet Corpus)[2]	Collection of several hundred documents giving instructions to patients about their medication.	WP4	
PubMed Central	200,000 full texts	WP1	
PubMed query logs	Set of 500 queries from PubMed logs		WP8
Radio2wiki	German radiology book	WP1, WP4	
Set of queries from General Public logs	Sets of query logs from the HON search engine: 1) set of 8181 queries collected between 24/11/2011 and 19/11/2011 2) set of 235,844 queries (EN, FR, ES) from December 2011 to February 2013		1) WP8 2) WP1
TRIP database logs[3]	Queries of 279,340 anonymous users from January 2011 to August 2012.		WP1, T1.5
Wikipedia	English wikipedia pages selected from the Health category	WP1	

[1] http://www.statmt.org/wmt11

[2] http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL/

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

B. List of image data sets used for training / evaluation

Data set	Description	Training	Evaluation
BioMedCentral image data set	Images from the freely available BioMedCentral journals and the free text	WP2	
DD1	Large scale realistic clinical data dump from MUW / AKH PACS.	WP2, WP9	WP2
FBCT1	Full body CT data	WP2	
HRCT images series with 3D annotated regions of pathological lung tissue	108 image series with annotated lung tissue patterns and clinical parameters. 85 image series with annotations used in the experiments and 5 lung tissue classes (of which 4 pathologic).	WP2	WP2, WP10
ImageCLEFmed 2009	12.677 (training) and 1.733 (test) 2D radiographs of various body regions	WP1	WP2
ImageCLEFmed 2010	Radiology captions (from articles published in Radiology and Radiographics - RSNA)	WP1, WP2, WP4	WP1
ImageCLEFmed 2011	231,000 images from PubMed Central	WP2, WP9	WP2
ImageCLEFmed 2012	Over 300,000 images of 75’000 articles of the biomedical open access literature (PubMedCentral)	WP2	WP2, WP10
L1	Lung images. It encompasses approx. 100 anonymized Lung CTs, with known anomaly.	WP2, WP9	WP2, WP9
Radio2wiki	Images from the German radiology book	WP2	
TH1	Lung CTs with contrast agent (aorta). Approx 30 keypoints, approx 20 patients	WP2	

C. List of terminologies, ontologies used for training / evaluation

Data set	Description	Training	Evaluation
FMA	Foundational Model of Anatomy	WP4	
ClinicalTrials.gov	Registry of federally and privately supported clinical trials conducted in the United States and around the world	WP4	
DBpedia	Large multi-domain ontology which has been derived from Wikipedia	WP4	
DrugBank [21]	The database contains nearly 4800 drug entries including > 1,350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and > 3,243 experimental drugs.	WP4	
Radlex	Ontology for radiology (English).	WP1, WP2, T2.2	
Radlex (German version)	Ontology for radiology (German)	WP2, T2.2	
Terminologica Anatomica	International standard on human anatomic terminology	WP1	
Terminology used internally at Vienna Hospital (copy)		WP1	
UMLS vocabularies	Very large, multi-purpose, and multi-lingual vocabulary database	WP1, WP4	

D. List of the major additional crawled data sources

The list below contains the 40 bigger additional websites crawled in terms of number of web pages. Only websites identified to complete the existing index of HON-certified websites are indicated here.

Web source	# pages	Language
PubMed / MEDLINE	2115230	EN
Abeceda lékaru	34526	CZ
English Wikipedia	33090	EN
Adresse Pharmacie	30115	FR
Apotheken in Deutschland	20722	DE
German Wikipedia	19509	DE
Czech Society for Oncology	17036	CZ
Bundesministerium für Gesundheit/Ministry of Health Austria	16617	DE
Food and agriculture organization of UN. Food safety and quality	16448	EN
Lékarské slovníky	16299	CZ
Diabetes UK	15209	EN
ClinicalTrial.gov	13860	EN
Docteurclic.com	13359	FR
SmartBrief	12857	EN
Tribuna lékaru a zdravotníku	12342	CZ
Department of health	10585	EN
Cochrane collection	10276	EN
Žena.cz	10048	CZ
Abeceda Zdravi	9950	CZ
Zdraví od A do Z	9845	CZ
Ordinace	9151	CZ
Vše pro vaše zdraví	8454	CZ
Krankenhaus Experte	8433	DE
Medknowledge - CHAD Score information	7684	EN
Improving emergency medicine patient care	7543	EN
Ministerio de Sanidad, Servicios Sociales e Igualdad	7491	ES
Institute for clinical and experimental medicine	7308	ES
Medical news today	6994	EN
Informacní server o zdraví	6872	CZ
IDF - International Diabetes Federation	6476	EN
Ministerio de Empleo y seguridad social	5980	ES
Lékárna.cz	5935	ES
Compedium Suisse des Medicaments	5914	FR
leitlinien.de	5880	DE
Vitalia.cz	5724	CZ
Health information and quality authority	5582	EN
Guidelines of the National Institute for Health and Clinical Excellence (NICE)	5350	EN
Deutsche Krankenhausgesellschaft (DKG)	4523	DE
Medisite	4521	FR
NHS direct	4474	EN