Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on consistency checking rules for information extraction

Deliverable number	*D5.4*
Dissemination level	*Public*
Delivery date	*31 May 2013*
Status	*Final*
Author(s)	*Konstatin Pentchev, Vassil Momtchev, Dimitris Markonis, Thomas Schlegl*

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

Abstract

In this deliverable we describe the application of the Khresmoi Knowledge Base to two radiology use cases – the modality classification of images and their retrieval. In order to achieve this we made use of an domain ontology and semantic information extraction techniques. A significant improvement in the modality classification was achieved by combining a novel semantic classification approach with the existing visual features method. Two advanced search services were developed that allow the generation of consensus terms from a set of reports and the retrieval of reports in which a set of terms co-occur.

Table of Contents

Table of Figures

List of Abbreviations

KB	Khresmoi Knowledge Base
TOS	Talend Open Studio
RDF	Resource Description Framework
KS	Large Scale Biomedical Knowledge Server
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
IE	Information Extraction
NER	Named Entity Recognition
SKOS	Simple Knowledge Organization System
URI	Universal Resource Identifier
SBT	Semantic Biomedical Tagger
MUW	Medical University Vienna
HES-SO	University of Applied Sciences Western Switzerland
UMLS	Unified Medical Language System

1 Executive summary

The consistency checking rules are a common approach to improve the data quality by detecting errors or conflicts in the generated information. In the scope of WP5, the Khresmoi KB deals with large datasets created as result of complex transformation or data generation jobs, which are persistently stored into an abstract information model – RDF. This deliverable explores the information semantics and links its usage to multiple project use cases by augmenting the capabilities of the visual retrieval system using background knowledge. The work behind this deliverable involves close collaboration between Ontotext, HES-SO and MUW.

Three new services were developed that make use of the semantic repository. The constraint service exposes an image modality classification algorithm. It makes use of information extracted from non-structured text associated with the images by linking the latter to instances from the RadLex ontology. A heuristic scoring scheme is applied to the extracted data to assign modalities to images. The approach on its own successfully assigns the imaging techniques to 82% of the test set. In addition, a combination of the semantic classification approach and the classification method developed by the HES-SO improves the correct assignment to 90%.

The other two services were developed in order to integrate semantic search into the clinical radiology retrieval platform. For this purpose data from unstructured radiology reports was extracted based on RadLex terms and loaded into the KB. Two advanced semantic search services were set up – one for calculating the consensus term set for a corpus of reports and one for retrieving reports in which a given terms co-occur. In combination, these two services and the visual retrieval system can augment an initial result list, based on visual features alone, with a summarization of important terms and additional results from reports whose semantics agree with this consensus.

Additional work is to be done in collaboration with USFD to improve the IE in order to derive more complete and comprehensive semantics from the unstructured sources. This is to allow more advances semantic searches including not only instances but relations between specific resources as well.

2 Introduction

The Large Scale Biomedical Knowledge Server (KS) developed into the context of WP5 is a semantic warehouse that integrates a large number of data sources. The knowledge base transforms all sources into graph data model (RDF) and partially resolves the semantic modelling incompatibility across the different datasets by applying light-weight semantics rules as described in D5.1 section 3.1 [1]. In deliverable D5.2 [2], we demonstrate how the resulting knowledge base is applied to several use cases. It is used as a database for entity lookups via the autocomplete service and input for a reasoner that infers new connections after the expansion of transitive, equivalent and symmetric ontology properties.

Task T5.4 puts the knowledge base use into a completely different approach. It will use the information semantics as a background understanding to verify that there are no logical clashes in two specific use cases based on heuristics.

2.1 Modality Classification of Radiology Images

Medical literature contains a huge amount of several types of images, diagnostic (e.g. radiology, dermatology, microscopy etc.) but also non-diagnostic (e.g. graphs, diagrams, photographs etc.). A recent survey on the radiologists' image search behavior showed that radiologists fail at about 25% of the cases when they search for radiology-related visual information using the current tools [3].

One of the most commonly requested functionalities by the radiologists when asked to describe an ideal image search system was the ability to filter the search by the image modality [3]. This was also validated by the outcomes of the first round of the user-centered evaluation of the Khresmoi system which are presented in deliverable D10.2. This functionality requires knowing apriori the type of all the database images, information which is not always available for article figures. Manual annotation is not feasible for very large datasets, so machine learning techniques were developed for automatic image modality classification of images using textual and visual information.

Content-based image retrieval (CBIR) systems [4] use low-level visual characteristics of the image (e.g. color, shape, texture) to find similar images in a database. CBIR was proposed to assist in medical image retrieval applications [5] and was shown to be improving the automatic image type categorization performance when combined with text information [6].

The Khresmoi 2D image retrieval system uses a combination of CBIR and text retrieval techniques to search into images of the medical literature. Modality classification can improve the system's image retrieval performance and efficiency by automatically filtering out non-diagnostic images and by giving the ability to narrow down the search among images of specific modalities.

2.1.1 ImageCLEF

ImageCLEF[1] is an evaluation challenge on cross language image retrieval tasks. One of the main tasks is that of medical image retrieval. In this task, a modality classification subtask is included where, each year, a common benchmark is set up to evaluate classification techniques based on textual and visual information.

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

In this study, the setup from ImageClef2012 was used [7]. From the full dataset of the 300,000 images, a subset of 2000 manually annotated images is split in two sets of equal size which are used as training and test sets. A state-of-the-art technique that combined textual and visual information achieved an overall accuracy of 68.6% [8]. The same approach was used in our experiments and achieved an accuracy of 78.8% on the subset of radiology images (which number 203). This was done in order to have a reference performance, which to compare with future improvements.

This level of classification performance may be sufficient for a classification between diagnostic and non-diagnostic images (reaching 91.6%), but is not sufficient to improve image retrieval by filtering using the modality, due to many misclassified images.

Several factors may be causing the low performance. Certain modalities may have the same low level visual characteristics (e.g. CT and MRI) and are often misclassified while not all of the journals control the quality of captions, and affecting this way classification using the text information.

A set of consistency check rules might be able to assist on decreasing the number of misclassified images and thus increasing the modality classification performance.

2.1.2 Semantics and Classification

In order to improve the precision of the modality classification, we would like to employ the structured knowledge from the KB. Our goal is to make use of the background knowledge available - that each modality is applied for the detection of a specific set of pathologies in specific anatomical regions and/or tissues. This constraint should allow us to use meta-data associated with the images in order to make predictions on the modality of the image.

Because we have no way of semantically analyzing visual features and no explicit meta-data is provided, we must make use of text associated with the images. This implies that an IE approach must be used in order to transform the unstructured knowledge from the text to structured meta-data associated with each image.

With regards to the specifics of the Khresmoi project and the semantic models available, we constrain our classification to radiology modalities only. Our hypothesis is that our approach will be able to correctly classify images to radiology modalities with high precision. Moreover, because we are going to use textual meta-data, the set of modalities we fail to correctly classify is going to be a different one than the set of incorrectly classified images by the HES-SO method, which makes use of visual features. Therefore, we plan to finally combine the two classification methods which should yield superior results than any of the two techniques on its own.

2.2 Radiology Image Retrieval

In some situations of daily clinical routine radiologists want to find cases that are similar to a given query case (volume). This particularly holds true for cases where the underlying image information shows pathologies or anatomical abnormalities that have a low probability of occurrence. Hence, the query is triggered on image information for which the radiologist searches similar cases. Whereas the query should return images and corresponding medical reports.

Furthermore, as the user tests showed, radiologists want to trigger literature search by a given image. The queried literature should contain images similar to the query image or should contain textual information about anatomy or pathology that is solely reflected by the underlying texture (visual appearance) in the query image.

Radiological images, acquired with modalities like MRI or CT often show similar intrasubjective appearances in different anatomical locations within the same organ (e.g. right lobe and left lobe of the

liver). On the other hand one can observe high intersubjective variations in the appearance of the same anatomical location of a given organ. Thus merely taking image information into consideration might be insufficient for the retrieval task.

A retrieval system that uses a combined model which integrates both image and semantic information might improve considerably the quality of the the results of the retrieval. The required semantic information can be extracted from radiological reports which comprise anatomy as well as pathology information/terms.

2.2.1 MUW Retrieval System

The current image retrieval system is solely based on image features. First we compute supervoxels (small regions that satisfy some conditions regarding homogeneity). Based on this supervoxels we compute local image features which reflect texture of the tissue in the local anatomical location. The actual query is triggered with and computed based on these local image features. So we do not search for cases which are entirely similar to the query case but we search for similar image batches within the query organ. Subsequently relevant cases are ranked based on the similarity of batches (supervoxels) as well as the number of top ranked batches within each volume. In the query case we do not have the corresponding radiological report. But we do have the reports of the query results and believe semantics can be used to breach this gap.

Calculating distances among reports and parts of reports could allow to compute a clustering of reports. Thus, we need a quantitative representation of reports and of parts of reports (terms) which furthermore allows us to define distance measures for reports and terms as well. This quantitative representation should also reflect the underlying hierarchical relations between the semantic tokens (e.g. single words or terms). This clustering would yield a set of top ranked reports (i.e. reports with minimal distances).

Additionally, using information semantics could yield consensus terms - under consideration of the anatomy-pathology relationships in the report - of the top ranked reports considering the fact that these terms are part of multiple interconnected hierarchies.

Based on these consensus terms the image retrieval system could additionally discriminate a ‘primary set’ and ‘differential diagnosis set’ and trigger corresponding literature search. Furthermore a comparison of the clustering in the reports and the clustering in the image domain potentially helps to distinguish how - if necessary - the appearance distance measures for image retrieval should be changed.

2.2.2 Semantic Information Retrieval

Semantic Information Retrieval, commonly referred to as Semantic Search, is about finding information that is not based on the presence of text (keywords, phrases), but rather on the meaning of the words. The problem with the keyword-based search engines is that, if this information is published by diverse sources, the same term may be used with different meaning and different terms may be used for concepts that have the same meaning.

Semantic Search engines try to bridge this gap by using semantics and thus offering the user more precise and relevant results. The approach takes advantage of conceptual models, such as ontologies, knowledge bases, thesauri, etc. [9] These models work at the human conceptual level, and at the same time they provide computer-usable definitions of the same concepts. By structuring the knowledge in a given domain, they offer common language that allows for more efficient communication and problem-solving.

In deliverable D5.2 [2] we reported on the creation of a knowledge base (KB) of structured information that will allow us to implement a semantic search platform for the radiology domain. By applying semantic annotation [9] on text linked with the images, we will extract valuable meta-data that is linked to resources from the KB. Using the named entities extracted from each report and their relations, we can then generate a feature space for each document and for a set of documents. This formal representation of the documents/images can be used to summarize, cluster and search.

3 Methods

In this section we describe the methods and approaches developed in order to achieve the goals of the two use-cases: image modality classification and radiology information retrieval.

3.1 Semantic Modality Classification

As stated in section 2.1.2, we would like to extract meta-data about the images from associated text and use this information to classify the modality to one of the categories described in section 2.1.1. In order to achieve this goal, we need a common model of the domain, to which both extracted data from free-text can be mapped and which describes the relationships between pathologies, anatomical regions and modalities.

In the current composition of the KB there is one source loaded, which covers all these requirements – RadLex. In the following subsections we describe the relevant parts of the RadLex ontology, give specifics on how it is used for IE and how we make use of its model to make predictions.

3.1.1 RadLex Ontology

RadLex is an ontology developed by the RSNA specifically to model the medical imaging domain. It contains a lexicon of terms, a hierarchy of these terms and relationships between the different classes of terms [10]. For modelling the ontology in RDF we used SKOS [11] semantics. Thus, every term is represented by an URI and has associated preferred and alternative labels, a type and broader/narrower relationships with transitive inference. For the purpose of this document we will look into the vertical and horizontal relationships of four classes of terms: imaging procedures and modalities, anatomical locations, pathophysiologies and observations. In the following list these four categories and their corresponding resource URIs are listed:

- anatomies
  - radlex:anatomy_metaclass
- pathologies
  - radlex:pathophysiology_metaclass
- modalities
  - radlex: RID10311
  - radlex: RID13060
- observations
  - imaging_sign_metaclass

The vertical relationships define the hierarchy of terms, e.g. super- and sub-classes of diseases, imaging procedures etc. The horizontal relationships represent inter-class dependencies and connections, e.g. pathologic conditions manifesting inn anatomy, anatomies and pathologies observed by specific modalities etc. (See Figure 1) It is exactly the radlex model of these relationships that we use as the foundation of our modality identification approach.

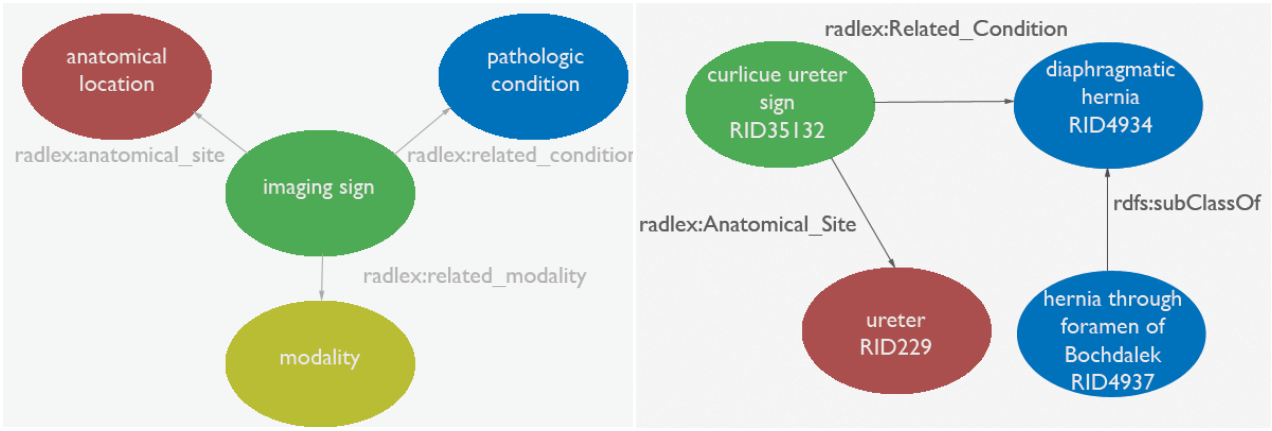

Figure 1: a) Inter-class (horizontal) relationships in RadLex. These links are between classes from different hierarchies. **b)** Example vertical and horizontal relationships. Vertical links describe a hierarchy of terms.

3.1.2 Semantic Information extraction

As the first step of our approach we need to extract structured information from the text associated with the images. To achieve this we aimed to perform NER of RadLex term of the previously described categories. The results of the IE will be made available for further analysis and use by enabling a view of the annotated documents and RDFizing the annotations into the KB (See Figure 2).

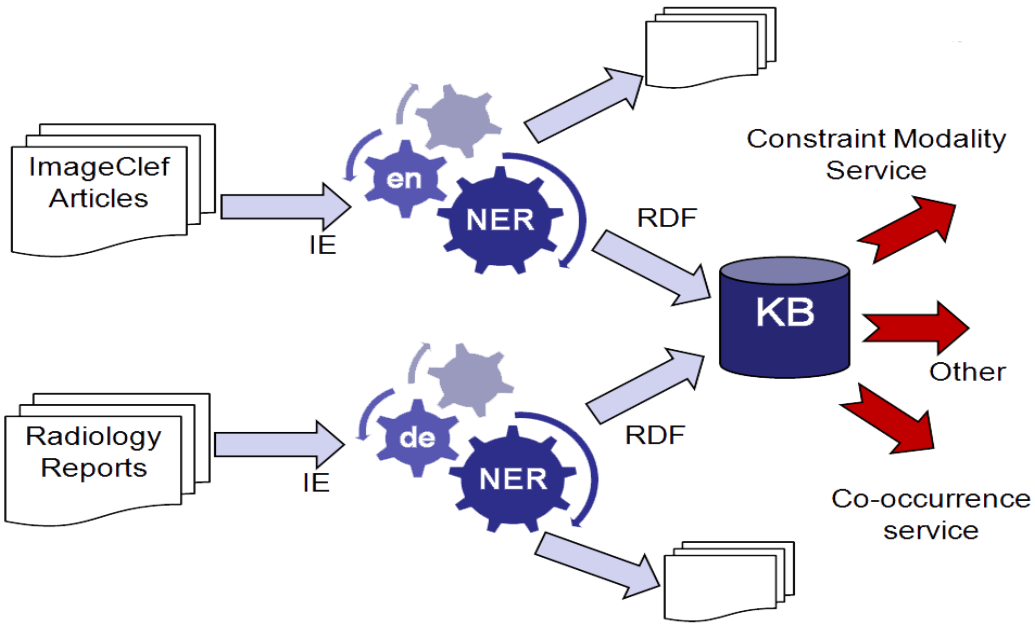

Figure 2 : Workflow of the IE pipeline and subsequent processing. ImageClef articles and radiology reports are processed by separate pipelines specifically designed for each task in english and german respectively.

The first issue we had to tackle when performing the NER was to associate text with images. It is common that a single article from the ImageClef corpus contains several figures that we wanted to

analyze. The only meta-information available to us was the caption associated with the image itself. We were presented with two choices – run the NER pipeline over only the captions or develop an innovative and very specific to this scenario approach, which sections the full text and associates sections with images based on their content. We decided for the former method, because it required one less step that might introduce noise and was guaranteed to contain only information relevant to the image, thus giving better precision to our approach. More details on the evaluation of the IE results are given in section 4.1.

In order to perform the IE we used GATE [12][13] as a platform and several resources that were previously developed on its basis. The pipeline uses a specific tokenizer, which rules are adapted to process English biomedical text. For example, the correctness of the chemical names finding required considering the comma between numbers and the hyphens as a part of the words. Then, OpenNLP POS tagger trained on GENIA corpus and Morphological analyzer are applied in order to discover the roots of the words. The roots are used together with the strings as token features on which the LD-Gazetteer performs NER based on a dictionary of resource URIs, labels and types. The LD-Gazetteer is developed for the SBT application [14], a commercial product offered by Ontotext. The dictionary can be populated using SPARQL queries (See appendix 7.1). All these additional processing steps improve the recall of terms, without having a negative impact on the precision.

The design of this NER allows us to link images to resources of the preselected types anatomies, pathologies, image findings and modalities. The results are all converted to RDF triples of the following scheme:

<http://khresmoi.ontotext.com/resource/imageClef/image/1477-7800-2-10-8> <lld:mentions> <radlex:RID2011>

In addition, each instance of the images is assigned an <rdf :type> <http://khresmoi.ontotext.com/resource/imageClef/image> , enabling us to easily distinguish them. For performing the batch processing we used TOS with the RDF and GATE components described in D5.1. The generated RDF is imported to the Khresmoi KB and can be easily accessed for further use using SPARQL.

3.1.3 Classification approach

The classification makes use of two facts:

1. The linking of images and RadLex resources through IE
2. The links between RadLex resources.

Indeed it appears that each pathology and anatomy is linked (through imaging signs) to a very specific set of modalities (See Figure 3). This allows us to use these links with confidence in order to associate a consistent set of modalities with an image.

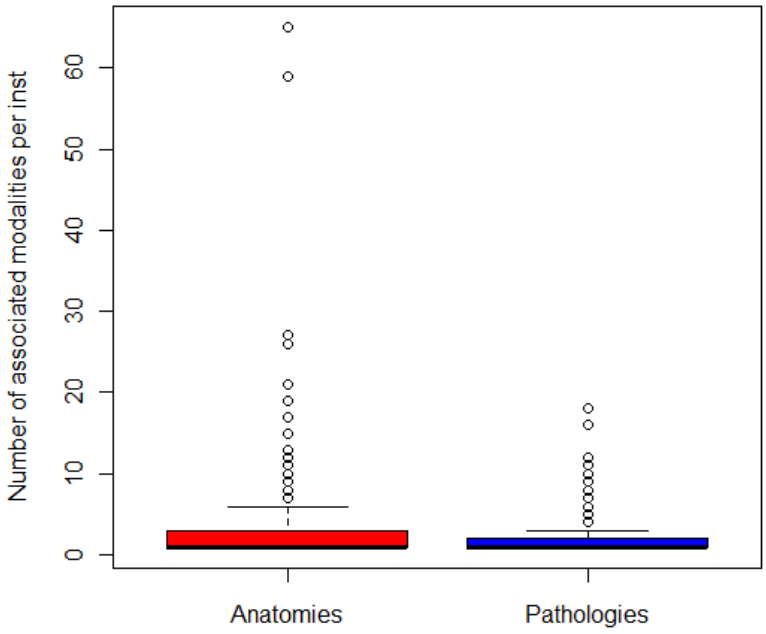

Figure 3 : A box-plot of the modalities per instance for anatomy and pathology classes. Both means are around 1.

The idea behind the approach is realized in that we use the RadLex terms associated with an image in order to assign it a modality in that we:

1. Assign a score of **1.0** for each term that is a *modality*
2. Expand terms of type *observation* to modalities and assign them a score of **0.5**
3. Expand terms of type *anatomy* and *pathology* to modalities (through observations) and assign them a score of **0.25**
4. Sum over all the scores of the modalities and sort descending; the first resource in the list is the predicted modality

The different scores are chosen in order to account for the confidence that can be derived from the mentioned term. Because an *observation* can be linked to more than one modality, we assign it only half the score of a mentioned *imaging modality*. Similarly, because *anatomies* and *pathologies* may be expanded to more than one observation, terms of these classes are given only half the score of an *observation*. The expansion of terms is implemented using SPARQL tuple queries (See appendix 7.3). In section 4 we discuss how the scoring can be improved for future use.

An additional improvement to this approach was implemented after close collaboration with HES-SO. A close analysis of the first results over the training set showed that often general concepts, e.g. *"Tomography"* (RID28840), were given the highest score, whereas the correct prediction should be a more specific resource, a sub-type, e.g. *"Computed tomography"*(RID10321). In order to compensate for this, an additional score modifier was introduced based on how far the resource is from the root of the hierarchy, i.e. how many resources are assigned as *rdfs:subClassOf* (with inference). For each link on the path from the root to the resource **0.2** is added to the score.

3.2 Semantic Image Retrieval

Similar to the approach described in section 3.1 we would like to extract meta-data about the images from associated text. The MUW has provided us with a corpus of 4500 anonymized radiology reports that are linked with radiology images/volumes from the Khresmoi system. Because the reports are in German, we had to develop a separate annotation pipeline, which again uses the RadLex ontology as its backing model. Otherwise, the process follows the depiction in Figure 2. We would then RDFize the extracted annotations. Two services are built upon the KB for searching through the reports – a co-occurrence search and a consensus search.

3.2.1 Semantic Information Extraction

A GATE processing resource was implemented that does basic NER on German text. For the task we again used the LD-gazetteer, but populated with term labels in German. The queries for populating the gazetteer cache are available in appendix 7.2. Again we extracted only terms of the modality, imaging sign, anatomy and pathology classes. We did not make use of any special analyzers or taggers, as we lack the expertise for doing information extraction on German text. The standard ANNIE tokenizer was employed [15], slightly adapted to bio-medical text by regarding dashes as non-separator symbols.

From the annotated text structured information is added to the KB with the following pattern:

<http://khresmoi.ontotext.com/resource/radiology/report/ra10001182381570> <lld:mentions> <radlex:RID10574>

The <rdf :type> assigned to the radiology reports is <http://khresmoi.ontotext.com/resource/radiology/report>.

3.2.2 Semantic Search Services

As previously mentioned we aimed to develop two additional services for the KS. The consensus service should allow deriving the set of resources – modalities, observations, anatomies and pathologies – common to a given set of radiology reports. This will satisfy the requirement from section 2.2.1 to retrieve consensus term spaces. The co-occurrence search is intended to yield reports in which a given set of terms co-occur, i.e. are all mentioned in the report. To implement both searches we required a system that can effectively perform facet queries that retrieve not only a result set but also calculate statistics for categories of values from the result set, e.g. in how many reports from a set the term 'hernia' is mentioned based on given search criteria. Because SPARQL queries are not efficient for such calculations we decided to implement an additional index over the terms using the popular IR engine Lucene/Solr [16], which offers capabilities such as:

- Faceted search and filtering
- Fast incremental indexing
- Multiple search indices

The approach that we implemented was to index the URIs of resources related to reports and then perform the faceted search using again URIs as input, basically making use not of the full-text search but only the indexing and faceted search capabilities. Lucene/Solr uses as its main data structure and inverted index but also keeps uninverted field copies, thus being optimized for both term and faceted queries.

3.2.2.1 Consensus search

The consensus search expects as input a set of radiology report URIs. This initial set will be generated by the image search described in section 2.2.1. The service then executes a faceted search for the union of term sets for each report. It delivers statistics for each term to how many resources from input set it is linked.

Example:

Input Reports	Modalities	Pathologies	Anatomies
ra10001175106360 **ra10001152715540**	Catheter angiography/RID10365 (2)	Ischemia/RID3376 (1)	Clivus/RID9332 (1)
ra10001174762610 **ra10001172760940**	Fluoroscopy/RID10361 (2)	Stenosis/RID5016 (2) Abcess/RID3711 (1) Lymphadenopathy/ RID3798 (1) Dilation/RID4743 (1) Fistula/RID4843 (1)	Saggital plane/RID10574 (2) Jejunum/RID148 (2) Fluid/RID1547 (2) Pelvis/RID2507 (2) Duod. Junction/RID32232 (1) Ileum/RID150 (1) Adrenal Gl./RID88 (1)

For a formal specification of the consensus web API see appendix 7.4.

3.2.2.2 Co-occurrence search

The co-occurrence search does the opposite of the consensus search – it retrieves reports for given terms. The terms are specified in a context identifying its significance for the report – is it a modality, pathology or anatomy. Therefore, the expected input is a set of field-terms pairs. The search returns documents whose term vector intersection for each context contains all provided terms. Because we again use the faceted search capability, not only are documents of interest returned, but statistics are calculated for terms outside the search input.

Example:

Input Terms	Reports
Pathologies: Emphysema (RID4799) **Anatomies: Spleen (RID86)** **Modalities: PET-CT(RID10341)**	8004900001173560 ra10001172849440 ra10001178377860

For a formal description of the co-occurrence web API see appendix 7.5.

4 Results and Evaluation

4.1 Semantic Modality Classification

For the evaluation of our classification method we processed the test set for ImageClef with the approached described in section 3.1. The corpus comprises of ~**70000** articles and **170464** images, which generated **338640** annotations. This corpus also includes the training set, for which we had the correct modalities. We wanted to compare the classification results for the training set to the 7 radiology categories. In order to achieve this, we first required a mapping between these categories and the RadLex resources which we used for classification. Such a mapping was already performed in [17], we give its summary below:

Table 1 : Mapping between ImageClef radiology categories and RadLex modality terms. Note that the last category – compound – is mapped to two terms.

ImageClef	RadLex	Label
DRUS	RID10326	Ultrasound
DRMR	RID10312	MRI
DRCT	RID10321	CT
DRXR	RID10345	X-Ray
DRAN	RID10371	Angiography
DRPE	RID10337	PET
DRCO	RID10341, RID10342	PET-CT, PET-MRI

We performed the classification evaluation by comparing the results of the semantic constraint method to the actual data for the 203 images from the training set. To better present the results of this comparison we plotted the confusion matrix given in Figure 4.

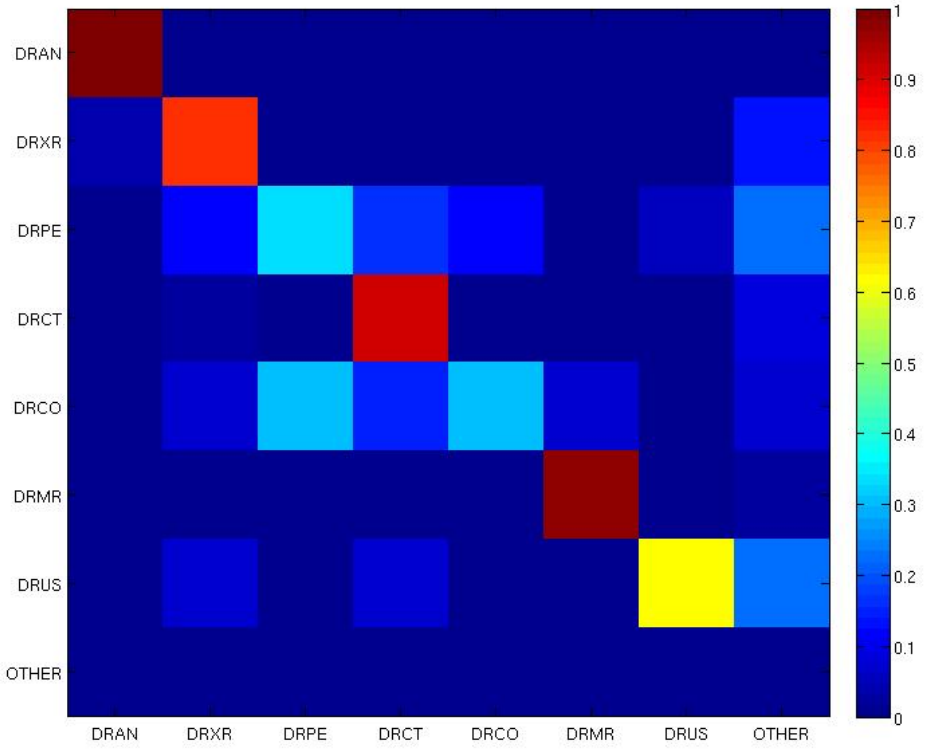

Figure 4 : The confusion matrix for the semantic classification approach. For each radiology imaging modality the percentage of assigned predicted modalities are plotted in a color-coded scheme.

The approach assigned the correct modality for 169 of the images - that is 82% success rate. While this is a good result that is higher than the 78.8% for the method described in [8], it is interesting to examine the categories for which the method underperforms. From the confusion matrix in Figure 4, we can derive that the semantic classification performs poorly for PET, PET-CT and PET-MRI. The ultrasound images are also classified correctly in only about 60%. Because we plan to use the two approaches together, it is interesting to examine which categories present a problem for the visual features method as well. The corresponding confusion matrix for the same set of images is given in Figure 5. It is visible that the method performs better for the PET, PET-CT and PET-MRI. However, it is less precise for Angiography and MRI. This fact raises the expectation that using the two methods in union will yield better results than either approach on its own. However, we were presented with the problem of how to combine the scores of the two methods. We experimented with the list of formulas for combining similarity values described in [18] and finally choose the *combSum* method:

$$S_i = w_a S_i^a + w_b S_i^b$$, where S is the score for an item i ; a and b denote different methods and w the corresponding weights

Using this approach of combining with weights 0.4 and 0.6 for the semantic and visual methods respectively, we classified 183 out of 203 images correctly. This is a success rate of 90.2 %, a huge improvement over the initial method which relied only on visual features. You can see the confusion matrix for these results in Figure 6.

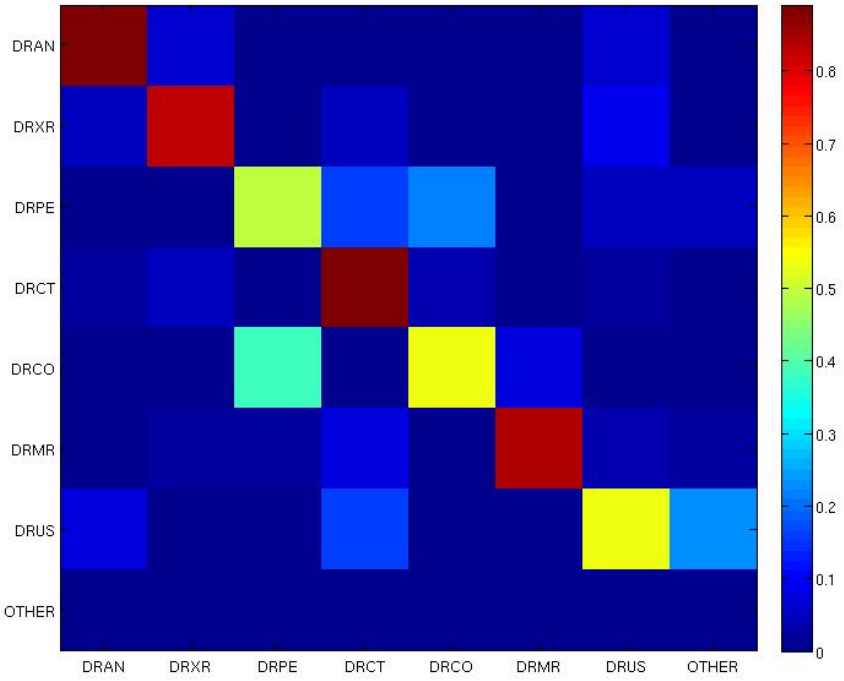

Figure 5 : The confusion matrix for the visual-features classification method. For each radiology imaging modality the percentage of assigned predicted modalities are plotted in a color-coded scheme.

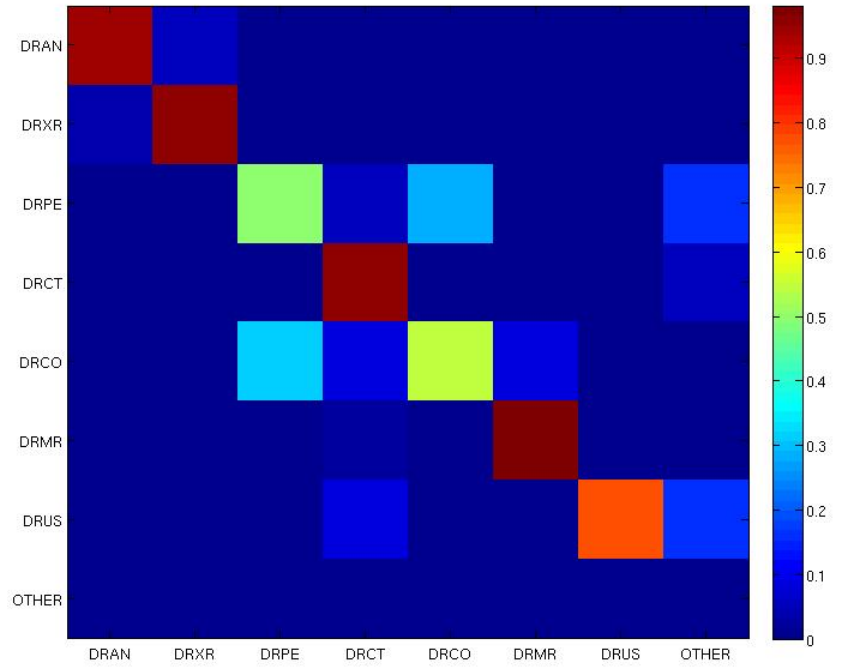

Figure 6 : The confusion matrix for the combination of the semantic and visual-features classification method. For each radiology imaging modality the percentage of assigned predicted modalities are plotted in a color-coded scheme.

4.2 Semantic Image Retrieval

The MUW provided us with a corpus of 5807 radiology reports. These were annotated and RDFized as described in section 3.2.1. We could extract links to RadLex resources for 4506 reports, the rest being either empty (372 documents) or not containing any of the gazetteer labels (929).

We developed an administrative interface, through which the user can specify which resources should be indexed for search and which linked resources are to be included as field indexes. The data is pulled out from the KB using user-defined SPARQL queries. For the radiology reports three fields were specified – modalities, pathologies and anatomies. The index is set up and can be tested on the KS at http://khresmoi.ontotext.com.

In addition, we also developed a web UI for the services so that these can be used by people (not only other services) and evaluated.

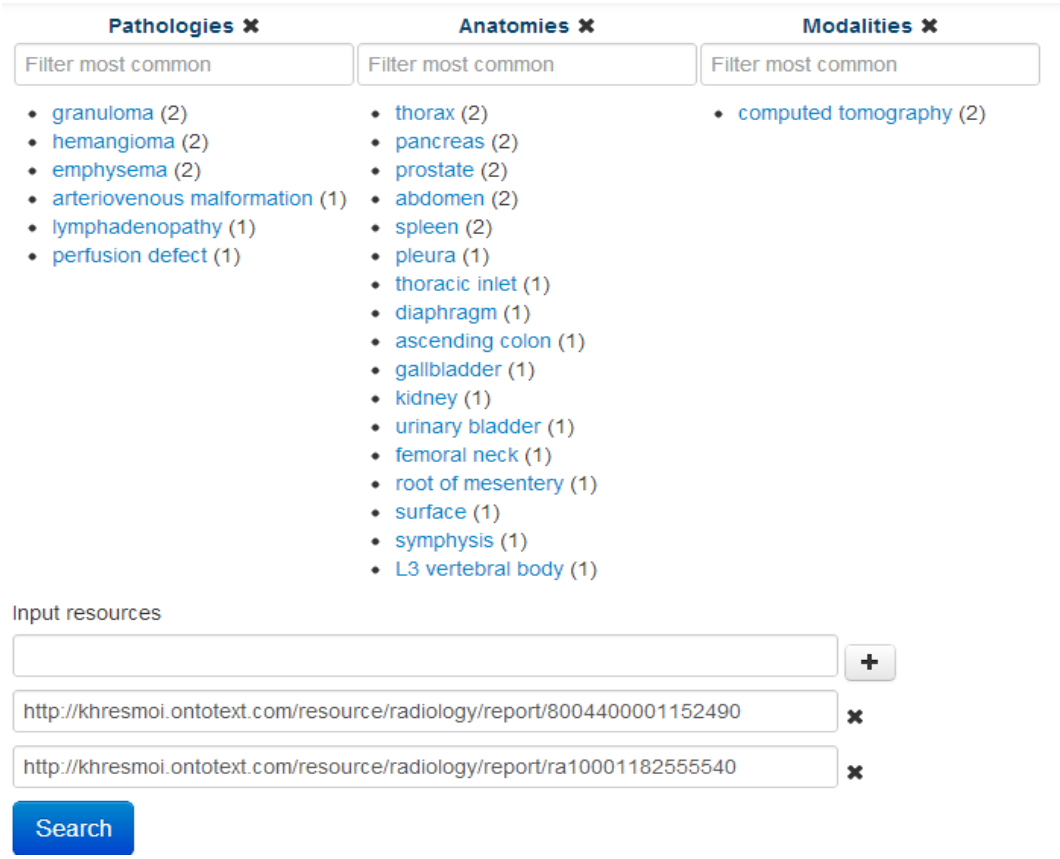

Figure 7 : The consensus search view, displaying results for two input reports: *8004400001152490* and *ra10001182555540.*

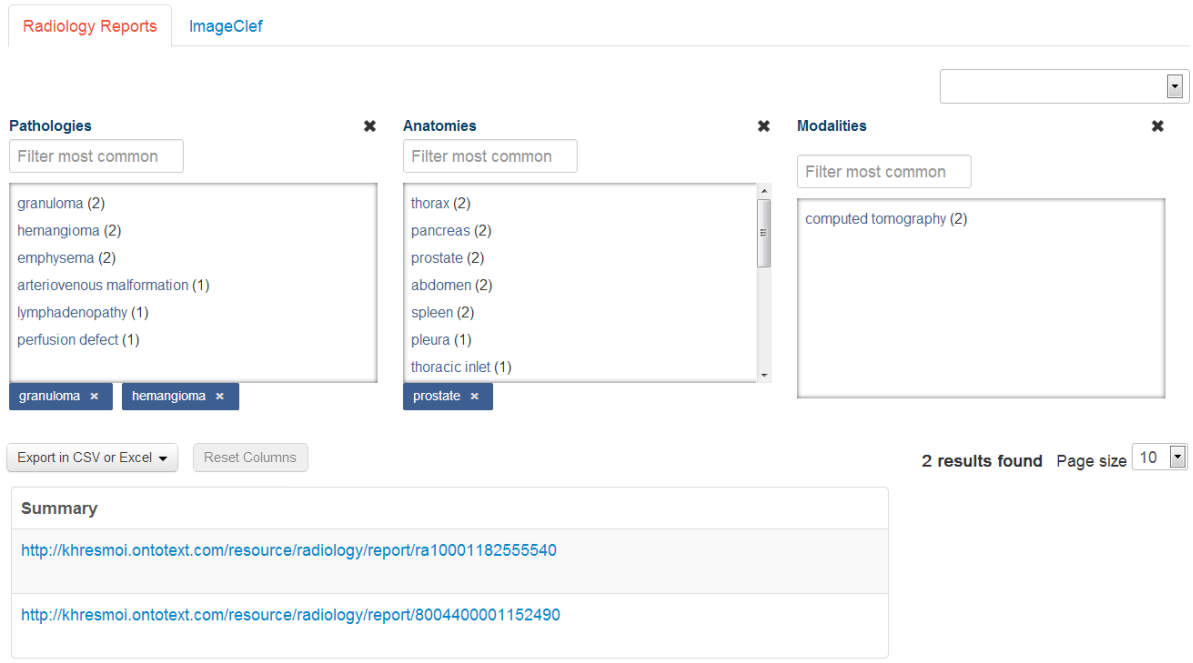

Figure 8 : The co-occurrence search view, displaying two results for a search for the pathology terms *granuloma, hemangioma* and the anatomy *prostate.*

In addition, the search was also set up for the ImageClef images per request by HES-SO and it is to be integrated with their search system. Both searches execute effectively, returning results in sub-second response times.

Further evaluation of the results returned in practice will be given in a separate deliverable by the use-case partners.

5 Discussion

Both the classification and semantic retrieval systems are by design dependent on the quality of the IE. The results of the work so far indeed showed one significant flaw – the recall of the annotations is low. There are a significant number of ImageClef image captions and radiology reports for which no RadLex resources could be identified. Manual analysis of the results of the IE showed that this issue is caused primarily by the quality of the gazetteer dictionaries used for NER. There are certain terms present in the unstructured sources for which no corresponding resources exist in RadLex. Our initial investigation of this matter shows that this is caused by an older version of the ontology present in the KB. We have already converted the newest version available to date (v3.8) to RDF and will pre-process all the input text accordingly.

However, a second cause of the problem will only partially be solved by updating the structured source. It appears that concepts are referred to in the text with labels (abbreviations and alternative names) that are not present in RadLex. This is especially true for the German radiology reports, as there is usually just a single German label available per resource. Even more, with assistance from the MUW we identified cases in the reports which make use of abbreviations that are institution specific. We have defined several steps in order to remedy these flaws. First and most important, the gazetteer caches are to make use of KB resources linked to instances from RadLex – namely UMLS concepts. The latter source contains rich IE information in terms of abbreviations, synonyms and is present in five languages [2]. A mapping of RadLex terms to UMLS concepts is already present in the KB. A second step we intend to take is to enrich the knowledge base with a restricted, manually compiled set of term labels specific to the radiology department of the MUW. While this approach is not a solution to the general problem, most of the compilation work was done during the manual assessment of the report annotation and will allow us to more precisely measure the impact of and the overhead required for having such a specific dictionary in the KB. Finally, the USFD is working on improving the tokenization of the German text, which as mentioned in section 3.2.1 was very basic. Negation detection is also an important improvement that we plan to make in order to improve the precision of the IE. Another important improvement to the IE is to perform not only NER over the radiology reports, but relation extraction as well. Currently, the use case is interested in the association of pathologies with anatomies in the context of a report. Knowledge of such relations will facilitate a more precise semantic search and consensus representation.

The last point for improvement with regard to the image modality classification is the scoring scheme we use. Currently, as described in section 3.1.3 we employ weights that were chosen arbitrarily after some experiments and worked well. However, this can be improved by using machine learning techniques in order set the scoring factors precisely. We did not do it at this stage, because from the incorrectly classified images only 3 had the correct modality scoring lower and the rest did not include the correct modality at all in the classification list. The latter issue is again linked to the IE problems discussed above.

6 References

[1] K. Pentchev, V. Momtchev. D5.1 Report on data source integration

[2] K. Pentchev, V. Momtchev. D5.2 Large Scale Biomedical Knowledge Server

[3] Dimitrios Markonis, Markus Holzer, Sebastian Dung, Alejandro Vargas, Georg Langs, Sascha Kriewel, and Henning Müller. A survey on visual information search behaviour and requirements of radiologists. Methods of Information in Medicine, 51(6):539–548, 2012.

[4] Dimitrios Markonis, Adrien Depeursinge, Ivan Eggel, Antonio Foncubierta-Rodrıguez and Henning Müller. Accessing the medical literature with content–based visual retrieval and text retrieval techniques. In Proceedings of the Radiological Society of North America (RSNA), November 2011.

[5] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine–clinical benefits and future directions.International Journal of Medical Informatics, 73(1):1–23, 2004

[6] Dimitrios Markonis, Alba Garcia Seco de Herrera, Ivan Eggel and Henning Müller. The medGIFT Group in ImageCLEFmed 2011. In Working notes of CLEF 2011. Amsterdam, The Netherlands, 2011.

[7] Henning Müller, Alba Garcia Seco de Herrera, Jayashree Kalpathy Cramer, Dina Demner Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classication tasks. In Working Notes of CLEF 2012 (Cross Language Evaluation Forum), September 2012.

[8] Alba Garcia Seco de Herrera, Dimitrios Markonis, Ivan Eggel, and Henning Müller. The medGIFT group in ImageCLEFmed 2012. In Working Notes of CLEF 2012, 2012.

[9] Kiryakov A., Davies J. Semantic Search . "Information Retrieval - Searching in the 21st Century"; Goker A. (Editor), Davies J. (Co-Editor), Graham M. (Co-Editor). John Wiley & Sons, Europe, 2007

[10] Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 26 (6): 1595-7. doi:10.1148/rg.266065168

[11] A. Miles, S. Bechhofer . SKOS Simple Knowledge Organization System Reference . W3C Recommendation. 18 August 2009

[12] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

[13] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.

[14] Ontotext, Semantic Biomedical Tagger, http://www.ontotext.com/semantic-biomedical-tagger

[15] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002

[16] The Apache Foundation. Apache Solr. http://lucene.apache.org/solr/

[17] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani . Creating a classification of image types in the medical literature for visual categorization . SPIE medical imaging, 2012

[18] E. A. Fox and J. A. Shaw, Combination of Multiple Searches, Second Text Retrieval Conference (Trec-2), 1994

7 Appendices

7.1 SPARQL queries for RadLex English labels gazetteer

7.1.1 Modalities

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>
SELECT ?concept ?type ?literal
WHERE {
{
  ?concept rdf:type ?type . ?concept rdfs:subClassOf ?superclass .
  FILTER(?superclass = radlex:RID10311 || ?superclass = radlex:RID13060) .
  ?concept radlex:Preferred_name ?literal .
  FILTER(?type = owl:Class)
}
UNION
{
  ?concept rdf:type ?type . ?concept rdfs:subClassOf ?superclass .
  FILTER(?superclass = radlex:RID10311 || ?superclass = radlex:RID13060) .
  ?concept radlex:Synonym ?literal .
  FILTER(?type = owl:Class)

}}
```

7.1.2 Pathologies, Anatomies and Imaging Signs

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>
SELECT ?concept ?type ?literal
WHERE {
{
  GRAPH <http://linkedlifedata.com/resource/radlex> {
    ?concept rdf:type ?type .
    ?concept radlex:Preferred_name ?literal .
    FILTER((?type != owl:Class) && (?type = radlex:anatomy_metaclass ||
?type = radlex:pathophysiology_metaclass || radlex:imaging_sign_metaclass))
  }
}
UNION
{
  GRAPH <http://linkedlifedata.com/resource/radlex> {
    ?concept rdf:type ?type .
    ?concept radlex:Synonym ?literal .
    FILTER((?type != owl:Class) && (?type = radlex:anatomy_metaclass ||
?type = radlex:pathophysiology_metaclass || radlex:imaging_sign_metaclass))
  }
}}
```

7.2 SPARQL queries for RadLex German labels gazetteer

7.2.1 Modalities

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>
SELECT ?concept ?type ?literal
WHERE {
{
  ?concept rdf:type ?type .
  ?concept rdfs:subClassOf ?superclass .
  FILTER(?superclass = radlex:RID10311 || ?superclass = radlex:RID13060) .
  ?concept rdfs:label ?literal .
  FILTER(?type = owl:Class && lang(?literal) = 'de')
}
UNION
{
  ?concept rdf:type ?type .
  ?concept rdfs:subClassOf ?superclass .
  FILTER(?superclass = radlex:RID10311 || ?superclass = radlex:RID13060) .
  ?concept radlex:Synonym ?literal .
  FILTER(?type = owl:Class)
}}
```

7.2.2 Pathologies, Anatomies and Imaging Signs

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>
SELECT ?concept ?type ?literal
WHERE {
{
  GRAPH <http://linkedlifedata.com/resource/radlex> {
    ?concept rdf:type ?type .
    ?concept rdfs:label ?literal .
    FILTER((?type != owl:Class) && (?type = radlex:anatomy_metaclass ||
?type = radlex:pathophysiology_metaclass || radlex:imaging_sign_metaclass)
&& (lang(?literal) = 'de'))
  }
}
UNION
{
  GRAPH <http://linkedlifedata.com/resource/radlex> {
    ?concept rdf:type ?type .
    ?concept radlex:Synonym ?literal .
    FILTER((?type != owl:Class) && (?type = radlex:anatomy_metaclass ||
?type = radlex:pathophysiology_metaclass || radlex:imaging_sign_metaclass))
  }

}}
```

7.3 SPARQL queries for expanding resources to modalities

7.3.1 Observations

```
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>

SELECT DISTINCT ?modality WHERE {
    ${inst} radlex:Related_modality ?modality .

}
```

7.3.2 Anatomies

```
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>

SELECT DISTINCT ?modality WHERE {
  ?observation radlex:Anatomical_Site ${inst} .
  ?observation radlex:Related_modality ?modality .

}
```

7.3.3 Pathologies

```
PREFIX radlex:
<http://bioontology.org/projects/ontologies/radlex/radlexOwlDlComponent#>

SELECT DISTINCT ?modality WHERE {
  ?observation radlex:Related_Condition ${inst} .
  ?observation radlex:Related_modality ?modality .

}
```

7.4 Consensus API

7.4.1 Request URI

/consensus(.format)

7.4.2 Parameters

Parameter	Required	Values	Example
q	yes	The q parameter is a composite value separated by double colons (\::) and double semicolons (\;;) . The syntax of the field is as follows: <fieldName>::<searchItem1>;;<searchItem> The syntax of a search item follows this format: <full-text search entry or URI>|<label> (important for web gui)	uri::http://.../resource/.../DB00983|Formotero l;; http://.../resource/.../7fe7c47d254...| Diehm
resultType	no	Type of index to look up	CSP, Document, Drug
cooccurrenceFacets	yes	A lsit of co-occurrence facets to use for search	Genes,Proteins,Drugs

7.4.3 Response

7.4.3.1 Formatrs

- CSV
- EXCEL
- XML

7.4.3.2 Response Model

- List<FacetField> facets

7.5 Co-occurrence API

7.5.1 Request URI

/cooccurrence(.format)

7.5.2 Parameters

Parameter	Required	Values	Example
q	yes	The q parameter is a composite value separated by double colons (\::) and double semicolons (\;;) . The syntax of the field is as follows: <fieldName>::<searchItem1>;;<searchItem> The syntax of a search item follows this format: <full-text search entry or URI>|<label> (important for web gui)	uri::http://.../resource/.../DB00983|Formoterol;; http://.../resource/.../7fe7c47d254...| Diehm
resultType	no	Type of index to look up	CSP, Document, Drug
columns	no	Comma-separated list of column names	Document_Type,Document_Title,Modification_Date,Study
count	no (10)	Size of returned result set	20
start	no (0)	Start offset	50
cooccurrenceFacets	yes	A lsit of co-occurrence facets to use for search	Genes,Proteins,Drugs

7.5.3 Response

7.5.3.1 Formatrs

- CSV
- EXCEL
- XML

7.5.3.2 Response Model

- List<FacetField> facets