Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on results of the WP1 first evaluation phase

Deliverable number	*D1.3*
Dissemination level	*Public*
Delivery date	*30 August 2012*
Status	*Final*
Author(s)	*Niraj Aswani, Liadh Kelly, Mark Greenwood, Angus Roberts, Matthias Samwald, Natalia Pletneva, Gareth Jones, Lorraine Goeuriot*

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

Abstract

In this deliverable we provide information on the manual annotation tasks that are being carried out by a team of professional annotators at a Khresmoi sub-contractor, Lighthouse, in the Philippines and managed by a team at The University of Sheffield in the UK. We provide information on a variety of topics related to these manual annotation tasks. Amongst other things, our main focus, in this deliverable, is on answering questions such as; what is the quality of the annotations produced by these annotators?, what sort of difficulties do these annotators face while annotating?, what have we done to address these difficulties?, did our steps make any difference? and how do these manual annotations compare with the annotations produced by an automatic application? The deliverable also includes the IR evaluation methodology which will be used to explore and develop retrieval techniques which will utilise the potential power of the rich document annotations.

List of Abbreviations

DC	Document Classification
EL	Entity Lookup
ELPS	Entity Lookup Pilot System
GATE	General Architecture of Text Engineering
HON	Health on the net
IAA	Inter Annotator Agreement
IE	Information Extraction
MIMIR	Multi-paradigm Indexing and Retrieval
MG4J	Managing Gigabytes for Java
PC	Paragraph Classification
PR	Processing Resource
TF-IDF	Term Frequency Inverse Document Frequency
UMLS	Unified Medical Language System

Table of Contents

1 Executive summary

A typical life cycle of a rule-based Information Extraction (IE) system begins with asking human annotators to annotate some sample documents. These annotated documents are then used as examples to develop a set of preliminary rules. These rules are then applied over a set of unseen documents. Mistakes are identified and used as feedback to improve the learned rules.

In this deliverable we provide information on the manual annotation tasks that are being carried out by a team of professional annotators at the Khresmoi sub-contractor Lighthouse[1] in the Philippines and managed by a team at The University of Sheffield[2] in the UK. We provide information on a variety of topics related to these manual annotation tasks. Amongst other things, our main focus, in this deliverable, is on answering questions such as; what is the quality of the annotations produced by these annotators?, what sort of difficulties do these annotators face while annotating?, what have we done to address these difficulties?, did our steps make any difference? and how do these manual annotations compare with the annotations produced by an automatic application?

We have previously provided details of earlier manual annotation tasks carried out by the same team in deliverable D1.1 [2]. In this deliverable we provide details of the manual annotation tasks carried out since the publication of that deliverable.

In this deliverable, we also include the IR evaluation methodology which will be used to explore and develop retrieval techniques which will utilise the potential power of the rich document annotations.

1 http://www.lighthouseip.com

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

2 Introduction

"What treatment should I be undergoing for the severe pain in my knees?"

Answering such a question, automatically, is a difficult job. But why is it difficult? If only there were a fixed set of health related questions, it would have been easy to create a database from which the answers to such questions could be retrieved. However, there are a virtually infinite number of health related issues which could form a question. The difficulty is in understanding such questions and interpreting them correctly. Natural Language Processing aims to understand and make sense of written text. Information Extraction (IE) is part of the process which, given a piece of text, aims to extract specific information such as names of diseases, treatments etc. to aid in answering complex questions such as the above.

A typical life cycle of a rule-based IE system begins with asking human annotators to annotate some sample documents. These annotated documents are then used as examples to develop a set of preliminary rules. These rules are then applied over a set of unseen documents. Mistakes are identified and used as feedback to improve the learned rules.

In this deliverable we provide information on the manual annotation tasks and try to answer a variety of questions related to these manual annotation tasks. For example,

- What are these tasks?
- How do annotators accomplish these tasks?
- What efforts do they have to put into these annotation exercises?
- What is the quality of annotations they produce?
- What sort of difficulties do these annotators face while annotating?
- How is the feedback collected?
- What are the steps taken to address these difficulties?
- Do these steps make any difference?
- How do these manual annotations compare with the annotations produced by an automatic application?
- What are the lessons to learn from these exercises?

In Sections 3, we attempt to answer the questions listed above.

Annotation is only the beginning of the story. The real power from annotating medical collections is the utility that can be made of these annotations in information need scenarios such as the one shown at the beginning of this section. The Khresmoi system will allow users to express these information needs through keyword-based queries, and through optional query refinement, filtering and categorising options. In Section 4, we provide details on the IR evaluation methodology which will be used to explore and develop retrieval techniques which will utilise the potential power of the rich document annotations we are constructing.

3 Manual annotation tasks

3.1 What have we done so far?

Earlier, in deliverable D1.1, we provided information on the annotation tasks related to annotating radiology reports, articles from Gene Home Reference and some work on document classification. These documents were annotated to identify mentions of anatomical terms and mentions of diseases. It was observed that the annotators had difficulties in annotating documents during the first few initial tasks. This was partly because the annotators had to learn new software that was being used for annotating documents and partly because the task as a whole was new to them.

Even though the guidelines for these tasks were produced very carefully, annotators were able to find anatomical and disease terms in the text for which the description of how to annotate them was missing from the guidelines. Even though the guidelines are produced by domain experts, it is very difficult to cover every single aspect of how different anatomical and disease terms may be expressed in online articles. Therefore, as described earlier, one of the main purposes of performing manual annotation tasks is also to discover new examples.

Feedback and comments provided by the annotators were used to update the guidelines and also to clarify the errors made by individual annotators. We also produced a couple of tools in GATE that can be used for calculating inter annotator agreements. The important outcomes of the manual annotations exercises, carried out by the Lighthouse team so far, can be summarised as below:

- When the annotators had started their first task of annotating anatomical terms, agreement between the annotators and a curator, who reviewed annotator's work and disagreements, was as low as 0.61. In other words, annotators had disagreed with the curator on 39 annotations out of every 100 annotations. However, since then the agreement figures have improved significantly to reach as high as 0.96 for one of the tasks carried out for annotating anatomical terms.
- The annotators have annotated a large number of documents from a variety of sources (e.g. MEDLINE, web documents). As a result, it has been possible to capture different ways of how anatomical and disease terms are mentioned in a variety of documents and enrich the annotation guidelines with the same.
- We can observe consistency in the performance of annotators -- both in terms of agreement and the speed with which they annotate documents.
- We have been able to identify annotators who have excelled at the task of annotation. Not only do these annotators produce high quality annotations but they also take less time to complete similar annotation tasks. In future, should we need more annotators annotations produced by such annotators can be given higher weight to derive consensus annotation sets automatically.
- We have been able to identify software glitches that the annotators had experienced while annotating various documents. We have also established effective ways of communicating with the overseas annotator team.
- As a result of all these annotation exercises, we have been able to get more than 1500 documents annotated with anatomy and diseases annotations.

3.2 Managing annotation tasks

Communicating with a team based on the other side of the globe brings its own challenges and every possible effort was made to improve communication between the two teams.

One of the biggest challenges identified during the first phase of the manual annotation tasks was to deal with the time differences the two countries have. The time difference between the United Kingdom and Philippines is 8 hours. This forced the two parties to communicate over emails. However, it only meant delays for the annotation team who cannot, due to the difference of time in working hours, talk to the Sheffield team at the time a query is raised. Even if they leave an email, they need to wait for the Sheffield team to respond. But by the time Sheffield team replies, it is often out of working hours for the Lighthouse team.

The annotation tool, Teamware[3], used by the team at Lighthouse allocates documents to be annotated to different annotators randomly. If the annotators have a query for a particular annotation, one choice that Teamware offers is that of saving the document for later reopening. However, the tool restricts annotators from proceeding to the next document until they have completed and finalised the current document. In other words, they cannot move forward to the next document unless and until they have finalised the document they are currently working on. In case a query has been sent to the Sheffield team, it means, they have to wait for a day for the correct answer to come, act accordingly and submit the document. The only other option is to submit the document with error(s). Even though such errors can be identified and rectified by curators[4], it certainly, adds unnecessary burden to the work of curators.

To avoid such problems, both teams adopted a new strategy. According to the new strategy:

- Only a handful of documents (between 10-50, depending on the length and types of documents) are selected randomly for the first round of annotations. The team at Lighthouse annotates these documents and collects any problems they may come across.
- Both the teams have started using a shared Google spreadsheet to keep track of problems that the Lighthouse team faces and also for recording the answers given by the members of the Sheffield team. This spreadsheet is shared among all the annotators participating in the annotation task so that they can also keep up with the questions posted by other members of the annotator team.
- Both the partners have agreed on organising an early stage annotation task (the first bullet point), at a mutually agreed time. The idea is for a person from Sheffield to be available during the annotation task and respond to any queries posted on the shared document (the second bullet point). If necessary, a Skype call is established to discuss any unanswered questions and a possible demonstration of annotations if there is a need for one.
- At the end of such an exercise, the Sheffield team uses these questions and answers to review the guidelines or take necessary steps to make the task easier for the annotators.

Below we provide a few example questions that were asked during one of the annotation task sessions where a member of the Sheffield team was also present. Here, the annotators were asked to go

[3] http://www.gate.ac.uk/teamware

[4] Curator is a person responsible for reviewing annotations created by the annotators and comes up with a consensus annotation set that contains only the correct annotations.

through a list of pre-annotated annotations (produced by using one of the IE applications developed in Sheffield), delete the incorrect ones and add any missing ones.

Question:
Can a profession be annotated as an organism?
Answer:
Please look at the following link: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html. The hierarchy presented on this page shows you relationships between the different semantic types. Your question can be rephrased as *Is profession appearing as one of the subtypes of organism?* If the answer is yes, you can annotate. If the answer is no, you must not. Now, if you look at the hierarchy, you will notice that it actually does not. "Professional" is appearing as a subtype of "Group" which is a subtype of "Conceptual Entity". Examples of subtypes of organisms are "Plant", "Virus", "Bacterium", "Bird", etc. I hope this is clear. "

Here, we give another example:

Question:
The guidelines say -- you should never add an annotation that overlaps one that is already present.
Answer:
Let's take an example: 'calf pain'. If there is no annotation on this string (not even on individual words), you will obviously want to annotate the entire string as a single annotation. However, once you do that, there is no need to annotate 'calf' and 'pain' as separate annotations as that would be overlapping new annotations with the longer 'calf pain' annotation.

If the person from Sheffield was not available when the query (as shown above) occurred, either the annotators would have annotated all professions as organisms or they would have had to wait until the next day when the question would have been answered. Similarly, the second question from the annotators suggests that there was a need for clarification in the guidelines.

Finally, we present one more example question.

Question:
Can we delete all annotations or lookups outside the section?
Answer:
Please do not bother to do so as that can be done automatically by a program.

In most cases, HTML pages have some header and footer information. Also they might have text available in the navigation panels to browse through the respective websites. It was decided that only the annotations that appear in the body of HTML pages should be preserved. Any annotations outside the area presented by the body annotations should be deleted. Thus, the team at Lighthouse was asked to concentrate only on the annotations within the body of HTML pages. If a member of the Sheffield team had not been available to answer this question, it is possible that the team might have wasted a lot of time deleting annotations outside the body area, which in reality could be easily deleted by a simple program as suggested in the answer. Also, it has been observed that the questions that came out during the question answering exercise were applicable in general and to various tasks.

After an initial exercise, the feedback is used for updating guidelines, and annotators are asked to annotate remaining documents using the new set of guidelines prepared based on the feedback. If after the initial exercise, a question arises which is not answered yet, the annotation team keeps track of

these special cases and continue to work on the other annotations of the same document. Once all the annotators have finished annotating all the documents, curators can open specific documents in Teamware and fix the errors. However, an important point to note here is that the frequency of such unseen questions seems to have reduced after using the pilot annotation tasks method.

3.3 Annotation tasks

As explained earlier in Section 3.1, one of the outcomes of the exercises carried out by the Lighthouse team was the discovery of annotators who are performing well and those delivering consistent annotations. These annotators were assigned three different types of tasks. In the first task, they were asked to assign each document a label such as "Biological risk", "Alcohol", "Sex" etc. depending on the content of these documents. In the second task, the annotators were asked to classify each paragraph and assign it a class and a subclass relevant to the content of the paragraph. An example of such a class is "Disease" and the relevant subclasses are "Definition", "Reasons, risk factors" etc. Finally, in the third task, the annotators were asked to annotate terms of the different types of concepts from UMLS (and not just the Anatomy and Disease).

Below we provide details of metrics used, followed by sections on the individual tasks.

Metrics and measurements

We measure performance of automatic systems against gold standards and against human correction of the system's output, for both annotation and classification tasks, using standard information retrieval metrics, as follows:

Precision is defined as:

$$P = \frac{true\ positives}{true\ positives + false\ positives} \quad \text{..... ...} \quad \text{eq. 1}$$

and measures the correctness in terms of what percentage of the annotations created by the system are correct, compared to the standard or the corrections.

Recall is defined as:

$$R = \frac{true\ positives}{true\ positives + false\ negatives} \quad \text{.....} \quad \text{eq. 2}$$

and measures the coverage of the system's results (what percentage did the system identify from all the items present in the standard or corrections).

The F statistic applied here is defined as the harmonic mean of P and R:

$$F = \frac{2PR}{P + R} \quad \text{.....} \quad \text{eq. 3}$$

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

For these calculations, 'positive' refers to an annotation in the human correction or standard, 'negative' refers to an annotation not in the human correction or standard, 'true positives' is the number of annotations produced by the system that are also in the standard or corrections, 'false positives' is the number of annotations produced by the system that are not in the standard or corrections, and 'false negatives' is the number of annotations in the corrections or standard that are not produced by the system. In defining positive and negative classes, we were lenient, in that we

allowed overlapping annotations to be considered matches. When considering an annotation feature, the annotations must match, and features must also match exactly.

Where a document is double annotated, we can define the agreement between annotators – this is known as the Inter Annotator Agreement, IAA. We use the method described in [5], summarised here. IAA is calculated from the number of matches and non-matches between the two annotators. For every match from the first annotator, there will also be a match from the second annotator. The total number of matches is therefore double the number of matches from any one annotator. The total number of non-matches is the sum of non-matches from each annotator.

IAA can then be calculated as:

$$IAA = matches / (matches + non\text{-}matches) \quad \text{eq. 4}$$

IAA can be shown to be equivalent to un-weighted F-measure [6].

3.3.1 Document classification

The purpose of the document classification task was to provide additional information for users searching documents that provide information concerning a specific thematic.

Annotators were asked to annotate only the first character of a document with a label specifying the class of the Document. In the guidelines[5], for each of the classes, a description explaining when a document should be classified into a specific category was provided. For example, for the "Biological risks" class, it was said that the document should be classified into that particular class only if the content discusses one of the following themes:

- Health hazards linked to exposure to bacteria, viruses, other microorganisms, fungi and associated toxins.
- Biological safety food.

Also, a couple of URLs linking to example documents with the classification were provided for each class. If a document cannot be classified into one of the provided classes, the annotators were asked to assign a special class, called "Unknown" to the document. When the annotators were not very sure about the selection of their classification, they were asked to leave a comment on the annotation.

It was observed that a document can have multiple topics mentioned in it. For example, an article published on the topic of alcohol may also discuss sex related issues. In such cases, the annotators were asked to add multiple class labels to documents.

Evaluation and Results

The Khresmoi partner Health on the Net (HON)[6] already had a collection of approximately 500 manually classified web documents, from previous work. This classified collection was used as a gold standard to examine the accuracy of annotations produced by the Lighthouse annotators. Details on these investigations are provided in this section.

[5] Guidelines of the document classification task and details of the first batch of document classification task are provided in the deliverable D1.1 [2].

[6] http://www.hon.ch

The annotators were given two sets of documents to classify. In the first batch, 72 documents and in the second batch 85 documents. All these documents were randomly selected from the collection of 500 HON classified documents.

First, the set of 72 documents was given to the Lighthouse team for classification, along with the guidelines and examples therein, mentioned at the beginning of this section. Each document was classified by three annotators and reviewed by a curator who decided the final document classification based on the classifications assigned by the annotators. Classified documents were then returned to the Sheffield team for review.

IAA figures were calculated between each annotator pair and between each individual annotator and the curator. As explained in the deliverable D1.1 [2], the Sheffield team has developed a tool that allows them to calculate IAA for both the entity annotation task and the document classification task. Output of this tool allows the Sheffield team to identify a pair of annotators and documents where the IAA is measured too high or too low. It also allows them to dig down to annotation level to see on which annotations the annotators had agreement or disagreement. In other words, to identify specific types of annotations or documents (depending on the task) where most annotators have disagreement.

	Batch 1	**Batch 2**
Number of documents	72	85
Average of IAA Macro averages	0.71	0.89
Average of IAA Micro averages	0.70	0.88
Average of Consensus Macro averages	0.74	0.81
Average of Consensus Micro averages	0.72	0.80
Precision	0.47	0.59
Recall	0.84	0.68
Number of documents with at least one correct class found	62	64
Number of documents with all the correct classes found	55	49

Table 1 Results of Document Classification Task

Annotations and classifications assigned by the annotators are compared to each other (see equation 4) and to other standards (see equations 1 & 2). Precision is calculated by dividing the number of correctly assigned labels to the documents (by the annotators) with the number of all the labels assigned to the documents (by the annotators). Recall is calculated by dividing number of correctly assigned labels to the documents (by the annotators) with the total number of labels present in the gold standard (supplied by the HON team).

The second column in Table 1 lists the IAA (eq. 4) and accuracy figures (eq. 1 & 2) for the first batch of the document classification task. Here, the IAA averages refer to the agreement between annotators whereas the consensus averages refer to the agreements between individual annotators and the consensus annotations. Also, we calculated the number of documents where the classifications assigned by the annotators have at least one class found for the respective documents in the gold standard. Similarly, we calculated the number of documents where all the classes assigned by the HON team were also assigned by the Lighthouse team in the respective documents.

At the end of the first batch, we found that the Lighthouse team were able to capture the right classification in most cases. This is visible in the recall of 0.84. However, there were many additional

classes added to documents which shouldn't have been there. This is reflected by the precision which is 0.47. In other words, more than double the number of classifications assigned by the HON team were assigned by the annotators. In the post annotation discussions it became clear that the annotators had assigned a class to a document even if they thought there was only a 50-50 chance that the classification was correct. Thus, the classifications can be actually assigned a confidence level. Following this, the annotators were asked to specify their confidence level when assigning a particular class to a document. The four confidence levels provided to them were "very confident", "confident", "possible" and "may be".

It was observed by the Sheffield team that the annotators were concentrating too much at the sentence level. In other words, they had provided classes for every possible issue discussed in a document, whereas the annotators should have been assigning classes based on the overall theme of a document. The task was not to read every sentence but to quickly skim through the document to get an overall idea of what the document is about. It was observed that often the title of the page gives annotators a hint about the topic of the document.

Given the aforementioned findings, the annotators were sent a second batch containing 85 documents with proper information on where the errors had occurred in the first batch and how to avoid them for the second batch. The third column in Table 1 lists the results of the second batch.

As can be seen in the results, although the annotators were able to reduce the number of classifications they assign to documents, the overall accuracy had reduced drastically (from 0.84 in the first batch to 0.68 in the second batch). It is interesting to note that the agreement level between all the annotators was very high which seems to suggest that may be there was something wrong with the gold standard. Following this, a meeting was organised with the HON team to discuss the classification results. It was found that the Lighthouse annotators were often assigning classifications that were relevant to a document's website as a whole, i.e. it's broader context, and not to the document itself. This may be in part because the annotators are being asked to consider the whole document – which includes general site headers and information – and not the core content. It is expected that sectioning the documents in order to split out this core content, and asking the annotators to concentrate on this, will improve agreement with the gold standard. This work will be carried through to year 3 of the project.

3.3.2 Paragraph classification

The original purpose of this task was to classify every individual sentence of a document. Annotators were asked to assign a class and a subclass to every sentence (classes are described in Appendix A of the guidelines of Sentence and Paragraph classification, Deliverable D1.1[2]). For every class and a subclass, information such as in which conditions those particular values should be used was provided. For example, for a class, "Disease", 12 subclasses, as shown below, were provided.

1. Definition
2. Reasons, risk factors
3. Mechanism of disease development
4. Symptoms
5. Diagnosis
6. Treatment
7. Prognosis
8. Associated conditions
9. Patient support

10. Day to day life
11. Clinical trials
12. Research and scientific publications

Given these subclasses, annotators were asked to annotate a sentence as "Definition" only if the sentence gives a definition of a disease. Similarly, the annotators were asked to assign "Reasons and Risk factors" as a subclass only if they find the sentence describing the reason a disease is caught or transmitted, or the risk factors for catching the disease. Similar instructions were provided for other subclasses as well.

When the annotators started annotating sentences, they found that there were quite a few problems with the task. For example:

- To define something, often in documents, more than one sentence was used.
- Classifying sentences did not make sense when the sentences were too short or incomplete. For example, name of an author given at the top of document or just a title that reads as "Symptoms". As sentence boundaries were recognised using an automatic IE application, this often led to incorrect sentence boundaries.
- Often it was observed that a class could be assigned to a sentence but not a subclass. For example, “Topical drugs". Here, it is only a kind of drug and there isn't any suitable subclass provided to say that. Should it be classified as a definition? But it could be also incorrect if the context of the sentence does not really define what the topical drugs are. In such cases, the annotators were asked to assign "Other" as the subclass.

Resulting from the problems observed with sentence classification, it was realised that classifying sentences did not make sense and therefore the annotators were asked to assign classifications to individual paragraphs. However, this introduced a new set of issues. For example,

- It was observed that it is often possible that for some paragraphs there was no suitable class available to be associated to it. For example a disclaimer at the end of a page. In such cases, annotators were asked to leave such paragraphs without any classification assigned to them.
- Since multiple annotators are asked to annotate the same documents, a potential issue with inconsistent paragraph boundaries was also identified. The annotators were requested to annotate such documents in two phases. In the first phase, the curators were asked to draw paragraph boundaries and then only distribute the documents among their annotators.

In Table 2, we provide IAA figures for the paragraph classification task. Since every paragraph was assigned a class and a subclass feature, we provide the IAA figures for class only feature and also for both the class and subclass features.

	Class feature	**Class and Subclass features**
Avg. consensus macro avg	0.84	0.79
Avg. consensus micro avg	0.81	0.75
Avg. IAA macro avg	0.81	0.76
Avg. IAA micro avg	0.79	0.72

Table 2 Inter annotator agreement for the paragraph classification task

It was observed that the users had better agreement on assigning a class feature. However the numbers

seem to be bit low for the subclass features. However, comparing classifications with the classifications produced by a curator, suggests reasonably high accuracy.

3.3.3 Entity lookup task

The core part of the Khresmoi annotation application creates semantic annotations i.e. annotations linked to concepts in the Khresmoi knowledge base. These annotations are used to assist in intelligent search and retrieval over the Khresmoi system. Clearly, the performance of the final system will therefore depend on the quality of these annotations. A methodology for improving the performance of the system has been developed that uses feedback from human annotators to correct system output. This is then fed into further development and improvement of the application. Annotators are given output from the system, and asked to correct it. Performance metrics of IAA, Precision, Recall, and F measure are used to assess the quality of the system, comparing system output to those of the corrected annotations. We must be careful not to consider the corrected output a gold standard, as the methods used to create it are not as rigorous as those used when defining a gold standard. The metrics do, however, provide useful information, and the final set of corrected output can form a standard, and even the basis of a gold standard.

In comparison to the Entity Lookup tasks listed in deliverable D1.1 [2], where the annotators were asked to annotate only the anatomy and disease annotations, here, the annotators were asked to consider mentions of all the semantic types supported by UMLS[7]. The task was to examine and correct annotations marked in consumer health information web pages, described as "Lookup" annotations as they are based on dictionary lookup, dictionaries being built from the Khresmoi knowledge base. The annotations are produced using an automatic IE application developed by the Sheffield team. The IE application, internally, uses the UMLS vocabularies to annotate the mentions with respective semantic types. More information on the application is provided later in Section 3.3.5 and Deliverable D1.2 [3].

Lookup annotations produced by the IE application contain a feature called "semanticType" that gives a general description of the annotation in text. "Body Location or Region", "Injury or Poisoning", "Disease" etc. are just a few examples of such semantic types[8]. As the annotators are asked to consider all the semantic types, they are presented with a significantly higher number of annotations in comparison to the previous exercises. However, for this task the annotators did not have to lookup every annotation in UMLS. They were asked to follow their knowledge of the field to decide whether the allocated semantic type was correct or not.

The HON team has crawled a large number of documents relating to the health information from the Internet. There are approximately 1.1 million English documents in this collection. A set of 500 English documents[9] was randomly chosen from this collection by the Sheffield team and sent to the Lighthouse team for the annotation task.

---

[7] "Unified Medical Language System (UMLS) integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records"

(http://www.nlm.nih.gov/research/umls/)

[8] http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[9] Please note that the 500 documents mentioned here are different from the 500 documents crawled by the HON team for the document classification task.

The annotators were asked to do two main things:

- Remove any incorrect annotation
- Add any obviously missing annotations, based on their own immediate knowledge.

In case of adding new annotations, the annotators did not have to specify a semantic type. They were asked to just highlight a span that was missed by the IE application and should have been annotated as one of the semantic types of UMLS. Since the annotators are dealing with a huge number of annotations with a variety of semantic types, the decision was made to keep the task simple.

3.3.3.1 Pilot annotation task

There were several observations made during the pilot annotation exercise:

- The documents were quite large in comparison to those the annotators had been annotating during the earlier exercises.
- There was a large increase in the number of annotations the annotators had to go through. On average there were approximately 1400 annotations per document. Hence, the time needed to annotate these documents was much higher than in the earlier tasks.

It was also observed that if certain types of annotations are excluded which are often ambiguous and not very interesting (i.e. from the perspective of Khresmoi processing or when it is likely that other more specific concepts would be present to cover the same spans), the number of annotations can be brought down significantly. Table 3 lists semantic types which were excluded.

Physical Object	Spatial Concept	Temporal Concept	Language	Classification
Organism	Geographic Area	Qualitative Concept	Organization	Finding
Anatomical Structure	Regulation or Law	Quantitative Concept	Group Attribute	
Manufactured Object	Organism Attribute	Functional Concept	Group	
Conceptual Entity	Intellectual Product	Idea or Concept	Substance	

Table 3 Semantic types excluded from the Entity Lookup Annotation task

Terms of such semantic types were automatically deleted from the documents, referred to as "class filtering" in the discussion below. As a result, the average number of annotations to be considered by the annotators came down to 729 annotations per document (i.e. reduction of 680+ annotations per document). The Sheffield team also found that if annotations outside the actual article content[10] are deleted, it can save a lot of the annotators' time. However, automatically identifying article content area is not an easy task and therefore the annotators were asked to manually go through the document and create an annotation covering the span of the actual article content – this span we refer to as a "Section" annotation, and we refer to excluding annotations outside of this span as "section filtering"

[10] Often HTML pages have text other than the actual article content. For example, advertisements, headers, footers, navigation panels such as menu items etc.

below. Deleting annotations outside the actual article content area further reduced the total number of annotations from 700+ to approximately 180 annotations per document. Table 4 provides statistics for the first 10 documents used in the pilot annotation task.

The first column in Table 4 lists the document numbers. The second column shows the number of annotations produced by the IE application for the document specified in the first column. Column 3 shows the number of annotations that were removed as a result of applying the class filter. In other words, annotations with semantic types as listed in Table 3 were deleted and the column 3 shows the number of annotations that were deleted. As specified earlier, the annotators were asked to create a Section annotation covering the span with the actual article content. Column 4 shows the number of annotations that were found outside the article content section. Such annotations were deleted as a result of applying the section filter. Finally, the last column shows the remaining annotations that are presented to the annotators for review. Thus, applying the class and section filters reduces the average number of annotations per document from 1411 to 180. Out of these 180 annotations per document, the annotators preserved approximately 80 annotations per document. In other words, they were actually deleting more than half of the annotations.

Document Number	# annotations produced by the IE application	# annotations removed as a result of applying class filter	# annotations removed as a result of applying section filter	# annotations presented to the annotators for review
1	1637	731	653	253
2	1336	610	682	44
3	1651	791	683	177
4	1579	760	644	175
5	1819	930	647	242
6	989	522	279	188
7	774	414	243	117
8	1595	733	693	169
9	747	391	278	78
10	1985	938	683	364
Total	**14112**	**6820**	**5485**	**1807**

Table 4 Annotation statistics for the Entity Lookup pilot study

Semantic type	% preserved annotations	% share in all preserved annotations	% deleted annotations
Occupational Activity	0	1.2	100
Natural Phenomenon or Process	0	1.1	100
...			
Pharmacologic Substance	49.5	5.9	50.5

Gene or Genome	50	0.2	50
Acquired Abnormality	50	0.1	50
Biologic Function	50	0.1	50
....			
Body Substance	89.2	7.3	10.8
Disease or Syndrome	92.1	5.7	7.9
Age Group	92.3	1.5	7.7
Cell Component	92.3	0.7	7.7
Biologically Active Substance	93.2	2.5	6.8
Family Group	94.7	1.1	5.3
Food	95.1	5.8	4.9
…			

Table 5 Numbers of preserved and deleted annotations during the Entity Lookup pilot study, presented semantic type wise

For the Entity Lookup task, the Sheffield team ran another experiment where they tried to find out the types of annotations which were deleted most by the annotators. These numbers were then compared with the number of annotations that were preserved for each of those semantic types. The idea was to see if it would be better to get rid of annotations of specific types which were found to be highly ambiguous, and on which the annotators had to spend a lot of time. It might be better to ask them to annotate terms of such semantic types if the number of terms of such semantic types is not too high. For example, see Table 5. Here, column 1 lists the semantic types; column 2 shows the number of annotations (of the semantic type presented in column 1) in percentages preserved by the annotators; column 3 gives the share of annotations (of the semantic type presented in column 1) in overall annotations (of all semantic types); and column 4 is the percentage of annotations deleted by annotators for the semantic type presented in column 1.

Given the fact that it takes more time to create a new annotation than just deciding whether the annotation should be deleted or kept, using these figures, it is possible to decide semantic types whose annotations should be deleted automatically. If all the annotations or majority of annotations of a specific type are deleted by the annotators, one can think of excluding such a semantic type entirely. For example, "Occupational Activity" or "Natural Phenomenon or Process". On the other hand, if a large number of annotations are preserved, such a semantic type must not be excluded. For example "Disease or Syndrome" or "Food".

The situation becomes tricky when the ratios are very similar. For example, see the figures in Table 5 for "Pharmacologic Substance" and "Acquired Abnormality". For both the semantic types, almost half of the annotations are preserved. But what is important to note here is that the former type has almost 6% share in the total number of annotations preserved whereas the latter has only 0.1%. In such cases, if it is decided to delete both the semantic types, the annotators will be spending a lot of time searching

for annotations of type "Pharmacologic Substance" whereas for the latter semantic type, it does not really matter.

As the analysis was conducted only on the first 10 documents annotated during the pilot study, the figures might not be very reliable and consistent across other documents. Hence, the decision was made to wait for more documents before filtering out annotations based on such statistics. When the details are obtained on a reasonable size document set, they will be further discussed in the future deliverables.

Is curation needed?

Each of these 10 documents was annotated by two annotators and finally curated by a third. When the annotations of different annotators were compared, it was found that the IAA figures were reasonably high. The overall figures for agreement between the annotators on these 10 documents were: 0.93 precision, 0.94 recall and 0.94 as f-measure. When the annotations produced by the annotators were compared with the annotations produced by the curator, it was noted that the f-measure was very high (i.e. 0.97) as well. This means, the annotators did not disagree too much among themselves and nor with the curator. Also, it should be noted that the annotators involved in these tasks have been annotating such documents for a long time now. Following this, the annotators were asked to skip the curation step completely. In the absence of a curator, all the annotations where an agreement could not be seen were automatically deleted.

Dealing with ambiguous annotations

While the annotators were busy annotating documents, an interesting observation was made. As suggested in [1], to combat the problem of ambiguous strings, any string that pointed to multiple concept IDs can be changed to point only to the concept with the lowest numerical value, with the intuition that lower concept IDs tend to refer to the common usage of a word or phrase. At the same time, our analysis of the 10 documents and a verbal communication with the annotator team revealed that that was indeed the case. Most of the annotations these annotators were deleting were overlapping ambiguous annotations.

The IE application implemented by the Sheffield team produces multiple annotations for a given string. For example, the string "bacteria" is annotated with two semantic types: "Functional Concept" where the id of the concept assigned to the string is C1510439 and "Bacterium" where the id of the concept assigned to string is C0004611. In such cases, the authors of [1] recommend preserving the concept with the lowest numerical value, i.e. "Bacterium". When counting the number of overlapping annotations, we found that on average, there are at least two annotations created for every single string in the document, which means, if ambiguous overlapping annotations can be removed using the approach proposed by King et al. (2011), we should be able to half the time annotators take to review existing annotations. Since the discovery was made after the annotators were already given documents to process, the process of filtering ambiguous annotations could not be implemented on these documents. However, the Sheffield team has added a relevant component (experimental) to their IE application to deal with ambiguous annotations and this will be used for all future annotation work.
After the pilot annotation task, the annotators' team had been sent three batches of documents to be annotated with the updated guidelines. In this section, we provide details of these three batches along with the further details on the documents used for the pilot study. As mentioned previously, documents in these batches were chosen randomly from the set of English documents crawled by the HON team. In Table 6, we provide details such as the number of documents, number of annotations produced by the IE application and number of annotations deleted due to the application of several

filters as discussed previously and finally the number of annotations preserved and added by the annotators.

Batch	# docs	#annots produced by the IE application	#annots deleted by the class filter	#annots deleted by the section filter	#annots presented for the review	#annots preserved by the annotators	#annots added by the annotators	IAA for preserved annotations	IAA for the new annots
1	10	14112	6820	5485	1807	690	23	0.93	0.74
2	130	223233	107705	70276	45252	19002	603	0.89	0.58
3	108	190216	91404	59304	39508	16792	811	0.92	0.71
4	100	165267	78553	53357	33357	14649	703	0.93	0.78
Micro average								**0.91**	**0.69**
Macro average								**0.92**	**0.70**

Table 6 Annotation statistics for the various Entity Lookup tasks

The IAA figures clearly suggest that the annotators have high agreement among them. Also as can be seen, the numbers of newly added annotations are very less in comparison to the overall annotations filtered out by the various filters and also in comparison to the number of annotations preserved by the annotators. This suggests that the filters introduced in the application are helping to improve the manual annotation task.

In the following section below, we list the number of person hours spent by the annotators to carry out various manual annotation tasks.

3.3.4 Person hours for various annotation tasks

In Table 7, we show the person hours spent on carrying out the manual annotation tasks. Here, the DC1 stands for the first document classification batch. Similarly the DC2 and PC1 refer to the second document classification batch and the first paragraph classification batch. ELPS is an acronym used for describing the pilot study for the Entity Lookup task Similarly EL in EL1, EL2, and EL3 refers to the Entity Lookup task.

As can be inferred from Table 7, the annotators had spent maximum time on annotating the first 10 documents (ELPS). This is due to the fact that no filters were used on the documents. It was only after the pilot annotation task that various filters were implemented and applied to reduce the number of annotations. It was observed that the annotators had difficulty in deleting overlapping annotations and this was a very time consuming process for them. Therefore, the annotators were given training and a demonstration of using Teamware shortcuts to speed up the process. The effect of this training can be clearly seen for Batches 3 and 4 in the Table 6.

Batch	#docs	#annotators	#annotators per document	#curators	Total annotation hours	#curation hours	Total hours
DC1	71	5	3	1	80	16	96
DC2	85	9	3	1	144	16	160
PC1	54	7	3	1	112	16	128

ELPS	10	5	2	1	84	84	168
EL1	131	5	2	-	501	-	501
EL2	108	5	2	-	167	-	167
EL3	100	5	2	-	167	-	167

Table 7 Person hours spent on carrying out the manual annotation tasks

In Section 3.3.5, we first provide a brief description of the IE application used by the Sheffield team to pre-process the documents. We then compare the annotations produced by the IE application with the manual annotations produced by the Lighthouse team.

3.3.5 IE application

The Sheffield team has developed an information extraction application to annotate terms such as names of diseases, mentions of human anatomical parts, treatments, symptoms etc. Further additions were made to the application to perform document and section classifications. This application was explained in detail in the deliverable D1.2 [3]. Here, we provide a brief overview of the application.

As explained in the deliverable D1.2, the application consists of several individual processing resources (PRs). These PRs help in producing semantic annotations. First, the application identifies word and sentence boundaries. Each individual token (word) is then assigned a grammatical category and a linguistically true base-form. The application uses an ontology-based gazetteer that annotates documents with UMLS concepts using the Khresmoi Large Scale Biomedical Knowledge base as described in the deliverable D5.2 [4]. Similar to this, a drug gazetteer is used for identifying names of many common drugs. The HON Tag parser is used for assigning HON classifications (based upon the domain from which the page originated) to each document. Also, the application has a PR to perform content detection. This allows the application to process and focus only on the actual article content. In order to assign labels to individual documents, the application processes individual sentences within the area of article content and uses a clustering technique over sentences to find the most probable labels. It uses a similar technique for section classification but relies on other resources as well: category and individuals gazetteers, ontology lookups etc. which provide hints for the labels to be assigned to individual sections.

Since, we have not been able to finalise the gold standards for document and section classifications (due to the reasons mentioned in Sections 3.3.1 and 3.3.2), in this section, we only concentrate on the evaluation of the entity lookup task.

In Table 8, we have compared the output of the IE application with the annotations produced by the Lighthouse team. Here we measure the accuracy of the IE application by considering the corrections produced by the Lighthouse team as the standard. Here, column 1 is the batch number. Column 2 is the number of documents in a batch. Column 3 is the number of annotations which were found to be common between the IE application and the annotations produced by the annotators. Column 4 gives the number of additional annotations which the annotators had created and IE application had missed. Column 5 gives the number of annotations which were deleted by the annotators or as a result of applying class and section filters. Column 6 gives the number of annotations on which there was a partial match (i.e. where the annotators had to fix boundaries of annotations produced by the IE application), Column 7 is the precision indicating how many annotations of the IE application are preserved by the annotators. Column 8 is the recall – i.e. if the figure is 0.98, it means that if the annotators had decided to finalise a document with 100 correct annotations, almost 98 annotations

were already identified by the IE application. Finally, Column 9 is the harmonic mean of precision and recall.

Batch number	# docs	Match	Only found by annotators	Produced by IE but deleted by annotators or filters	Partial match of span	Precision	Recall	F-Measure
1	10	706	7	1101	0	0.39	0.99	0.56
2	130	19313	279	25926	13	0.43	0.99	0.60
3	108	17280	277	22182	46	0.44	0.98	0.61
4	100	15054	294	18299	4	0.45	0.98	0.62
Total	**348**	**52352**	**857**	**67508**	**63**			
Micro averages						**0.44**	**0.98**	**0.61**
Macro averages						**0.43**	**0.99**	**0.60**

Table 8 Performance of the IE application when compared to the manually corrected annotations produced and preserved by the human annotators

Here, the recall suggests that the application has a wide coverage (almost 98 annotations out of 100 annotations preserved by the annotators are correctly identified by the application). However, the average precision figure suggests that efforts need to be spent on introducing further filters to reduce the number of annotations produced by the IE application. As specified earlier, the ambiguity filter is expected to reduce the number of annotations by half. However it would be interesting to see its effect on the precision and recall figures presented in the table above. We are hoping to increase the precision figure without impacting on recall.

4 Information retrieval evaluation methodology

Retrieving the required information in response to users' queries is a key requirement of the Khresmoi system. Various information retrieval (IR) approaches which harness the potential power of the rich annotations available in the Khresmoi document indexes will begin to be explored in the coming months. Exploring, developing and evaluating retrieval techniques necessitate a test collection consisting of indexed domain specific documents (document set), domain specific queries (query set) and lists of documents deemed relevant to the queries (relevance set). Our evaluation document set is the MIMIR indexed documents, described in D1.2 [3]. Section 4.1 describes the query set for the current evaluations of retrieval in Khresmoi. The associated relevance set is described in Section 4.2. Following this we provide an overview of the form of the evaluations that will be conducted.

4.1 Query set generation

A set of 100 English queries has been generated for evaluation purposes. This set consists of 50 'real' general public and 50 'real' general practitioner queries. Details on how the queries were selected follow.

The 50 English general public queries were manually selected from a sample of raw queries from the HON search engine[11] collected over a period of 6 months. Only non-capitalized queries were taken into account to remove possible influence by web crawlers using predetermined queries. The 50 queries were selected by a domain expert (Natalia Pletneva / HON). Queries which seemed to be too "medical" (for example, complex medical terms) and in languages other than English were excluded.

The 50 English general practitioner queries were manually selected by domain experts (Matthias Samwald / TUW, Marlene Kritz / GAW) to represent a variety of common search phrases found in the available query logs (PubMed[12] query log, Trip database[13] query log). Queries that contained spelling mistakes or which seemed not to stem from clinical information needs of medical professionals were excluded.

4.1.1 Query set analysis

Analysing the general public query set, we found that the queries are very short, broad queries (38 queries with one term, 12 queries with two terms), e.g. query = 'diabetes', query = 'eating disorders'. Given queries of this nature, we have no way of knowing what the information needs of the individuals who entered the queries were. We speculate that the possible information need scenarios for queries of this nature most likely were either: 1) the searcher does not want very specific results, but rather a collection of information including diagnosis, symptoms, lifestyle, etc; or 2) the searcher is being rather lazy and relying on the search engine to find something relevant to a more specific need or lacks the knowledge to enter a more specific query.

[11] http://www.healthonnet.org/

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

[13] www.tripdatabase.com/

The queries in the general practitioner query set contain a mixture of broad, short queries, e.g. query = 'myeloma', and more specific targeted queries, e.g. query = 'retinal macular degeneration, antioxidant treatment'. Here the average query length is 2.74 and the query length range is 4. We also note that the general practitioners show greater query generation knowledge, through their use of Boolean and clause quotes in queries, e.g. query = 'angiogram AND patient education', query = 'diabetes drug "weight gain"'. As for the general public, the information need for the broad queries are speculated to be either: 1) looking for a collection of information on the topic; or 2) looking for specific information, which they either were too lazy to attempt to specify in the query or were incapable of specifying in the query.

4.2 Relevance set generation

As is standard in IR evaluation, pooled result sets will be created for the 100 queries (described in the previous section) for relevance assessment. As part of this process queries will be annotated with LinkedLifeData[14] lookups. The pooled result sets will be generated using the MIMIR MG4J BM25Scorer and TfIdfScorer[15] ranking models on the text of the queries, on the lookups generated for the queries only, and on both the text of the queries and generated lookups in combination.

Relevance assessments will be conducted on the items in the pooled result sets by Lighthouse using provided software. Given the nature of the queries and the different possibilities for information needs related to these queries (discussed in the previous section), standard Boolean relevant/irrelevant relevance assessment will not suffice. We currently envisage Lighthouse labelling documents as relevant to different aspects/facets relating to the general queries, e.g. symptoms, treatment, diagnosis.

4.3 Evaluation approach

The generated indexes, queries and relevance sets, described in the previous sections, will be used for retrieval technique development, tuning and evaluation. Our evaluations will include, investigating the utility of the retrieval algorithms available in MIMIR MG4J (CountScorer, BM25Scorer, TfIdfScorer and HitScorer)[16] on the text based queries and on the LinkedLifeLookups we will annotate queries with. Various disambiguations of these approaches will also be investigated, including the possible utility of using other MIMIR annotations in the retrieval process.

As part of these evaluations we will include meeting the possible information needs underlying the queries (discussed in Section 4.1.1). Having Lighthouse label documents as relevant to different aspects/facets relating to queries, mentioned in the previous section, would allow us to study differences in retrieval behaviour for different aspects as elements of the system are changed, and also to calculate the effective retrieval effectiveness for a user interested in only one aspect, but who enters a more general query.

14 http://linkedlifedata.com/

15 http://gate.ac.uk/mimir/doc/mimir-guide.pdf

16 http://gate.ac.uk/mimir/doc/mimir-guide.pdf

5 Conclusions

In this deliverable, we have presented details of the three different tasks of manual annotations carried out by a team of annotators at Lighthouse. Our main focus in the deliverable has been to answer the questions: what is the quality of annotations produced by these annotators?, what sort of difficulties these annotators face while annotating?, what have we done to address these difficulties?, did our steps make any difference? and how do these manual annotations compare with the annotations produced by an automatic application?

We presented various IAA figures for different annotation tasks. In case of the document classification task (Section 3.3.1), where the external gold standard was made available, we compared the annotations produced by the annotators with the external gold standard. Even though the figures obtained are not as high as those obtained for other tasks, the entire exercise has been helpful in identifying problems related to the document classification task. One of the findings was that it is absolutely necessary to be able to identify parts of documents that should not be considered when classifying a document. It is also possible that multiple labels are applicable on different documents. Findings such as these can certainly help in designing an automatic document classification application as suggested in the section on the IE application (see Section 3.3.5).

With the paragraph classification task (Section 3.3.2), which was set out to be a sentence classification task in the first place, we found that it is very difficult to assign a classification to individual sentences. We provided details of the problems encountered and the reasons for considering paragraphs for classifications. Similar to the document classification task, the findings of this task were also used by the IE application developer to come up with a strategy to classify individual paragraphs.

In case of the entity lookup task (Section 3.3.3), we discovered that there were, simply, too many annotations produced by the IE application. We showed how a pilot annotation task helped us to develop various filters that reduced the number of annotations from 1400 annotations per document to 180 annotations per document. Our comparison of these annotations with those preserved by the annotators clearly suggests that the filters are indeed improving performance. Also the small number of newly added annotations by the annotators is a clear indication that by introducing such filters we are not removing many useful annotations.

The various strategies developed so far to communicate with the annotators and managing the annotations tasks have helped speed up the execution of annotations tasks.

We also described, in Section 4, the IR evaluation methodology which will be used to explore and develop retrieval techniques which will utilise the potential power of the rich document annotations we are constructing. These IR evaluations will form part of the larger global empirical IR evaluations, in which we will test the effect of different components of the system on each other. More precisely these evaluations will test different retrieval techniques as described in Section 4, but will also look at the impact other components have on retrieval performance, e.g. how does the ability to correct spelling errors impact retrieval, how does the ability to translate queries from German to English impact retrieval. The IR evaluations will be subsumed in the global IR evaluation, conducted in Autumn 2012, and reported in Khresmoi deliverable D7.2.

6 References

[1] King, B., Wang, L., et al. (2011). Cenagage Learning at TREC 2011 Medical Track. The Twentieth Text Retrieval Conference Proceedings (TREC 2011), Gaithersburg, MD. National Institute for Standards and Technology.

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

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

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

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

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