# CLEF-IP 2012: Retrieval Experiments in the Intellectual Property Domain

Florina Piroi[1], Mihai Lupu[1], Allan Hanbury[1],
Walid Magdy[2], Alan P. Sexton[3], Igor Filippov[4]

[1]Vienna University of Technology,
Institute of Software Technology and Interactive Systems,
Favoritenstrasse 9-11, 1040 Vienna, Austria
[2]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
[3]University of Birmingham, School of Computer Science,
Edgbaston Birmingham, B15 2TT, United Kingdom
[4]Chemical Biology Laboratory, SAIC-Frederick, Inc.,
Frederick National Lab, Frederick, Maryland, 21702, USA

**Abstract.** The CLEF-IP test collection was first made available in 2009 to support research in IR methods in the intellectual property domain. Since then several kinds of tasks, reflecting various specific parts of patent expert's work flows, have been organized. We give here an overview of the tasks, topics, assessments and evaluations of the CLEF-IP 2012 lab.

## 1 Introduction

The patent system encourages innovation by giving an advantage to people and/or companies that disclose their inventions to the open society. The advantage consists of exclusive rights on the published invention for a limited period of time, usually 20 years. A requirement for obtaining a patent, i.e. exclusive implementation and utilization rights, for an invention is that no similar invention was previously disclosed. Because of the high economic impact of a granted patent it is important that the specific searches during the examination of patent applications are thorough.

Current rates of technological development has resulted in a large increase in the number of patent applications filed with the patent offices around the world. To keep up with the increase in the amount of data, appropriate information retrieval (IR) methods have to be developed and tested. The CLEF-IP test collection gives the IR tool developers a test bed to evaluate the performance of their tools in the intellectual property area (specifically patents).

Since 2009 the CLEF-IP evaluation campaign (2009) and benchmarking labs (2010 and 2011) have posed their participants tasks that reflect specific parts of a patent examiner's work-flow: finding prior art of a patent application, classifying the patent application according to the International Patent Classification System, using images in patent searches, classifying images occurring in patents.

This year there three tasks were organized, each concerning a different aspect of the data that can be found in a patent collection:

– Passage retrieval starting from Claims. The topics in this task were claims in patent application documents. Given a claim or a set of claims, the participants were asked to retrieve relevant documents in the collection and mark out the relevant passages in these documents.
– Flowchart Recognition Task. The topics in this task are patent images representing flow-charts. Participants in this task were asked to extract the information in these images and return it in a predefined textual format.
– Chemical Structure Recognition Task. The topics in this task were patent pages in TIFF format. Participants had to identify the location of the chemical structures depicted on these pages and, for a specified subset of those diagrams, return the corresponding structure in a MOL file (a chemical structure file format).

The ideas behind the CLEF-IP task definition and their organization are instances of the use cases defined in the frame of the PROMISE[1] project.

The rest of the paper is organized as follows: Section 2 describes the CLEF-IP collection and each of the organized tasks. We detail, for each of the tasks, the idea behind proposing such a task, the sets of topics and their relevance judgements, and the measures used in assessing the retrieval effectiveness. Section 3 presents the results of the evaluation activites for each of the three tasks and a list of participants. We finish with Section 4.

## 2 The 2012 CLEF-IP Collection

This year's collection corpus is the same as the one used in the CLEF-IP 2011 lab, i.e. it contains patent documents derived from European Patent Office (EPO) and World Intellectual Property Organization (WIPO) sources, stored as XML files, corresponding to over 1.5 million patents published until 2002. This year's collection does not include the image data used in the Image classification and Image Retrieval task in the 2011 lab [6] . For a detailed description of the CLEF-IP collection we refer the reader to the previous CLEF-IP overview notes [6,7,8]. For a description of key terms and steps in a patent's life–cycle see [5]. To make this paper self contained, we re-state some facts about the CLEF-IP collection.

In the process of patenting, several, and potentially many, documents are published by a patent office. The most common ones are the *patent application*, the *search report*, and the *granted patent*. In most of the cases when a patent application is published, the search report is contained in it, otherwise it is published at a later date. Each patent office has its own identification system to uniquely distinguish the different patents. At the EPO and WIPO all patent documents belonging to the same patent are assigned the same numerical identifier. The different types of documents (application, granted patent, additional search reports, etc.) are distinguished by *kind codes* appended to the numerical identifier. For example, EP-0402531-A1 is the identifier of the patent application

[1] http://www.promise-noe.eu

document (kind code A1) of the European patent number 0402531, while EP-0402531-B1 identifies the European patent specification (i.e. text of the granted patent)[2].

The EP and WO documents in the CLEF-IP collection are stored in an XML format and are extracted from the MAREC data collection. MAREC contains over 19 million patent documents published by the EPO, WIPO, US Patents and Trade Organization and the Japan Patent Office, storing them in an unified XML format under one data definition document.

All XML documents in the CLEF-IP collection, according to the common DTD, contain the following main XML fields: bibliographic data, abstract, description, and claims. Not all XML documents actually have content in these fields. For example, an A4 kind document is a supplementary search report and will therefore not usually contain the abstract, claims and description fields. The documents have a document language assigned to them (English, German or French). The main XML fields can also have one of the three languages assigned to them, which can be different from the document language. Some XML fields can occur more than once with different language attributes attached to them. For example, EP patent specification documents (EP-nnnnnnn-B1.xml) must contain claims in three languages (English, German and French).

We continue this section with a more detailed description of the three tasks organized in CLEF-IP 2012.

### 2.1 Passage Retrieval Starting From Claims

The importance of the claims in a patent is given by their role in defining the extent of the rights protection an applicant has secured. The decisions taken by patent examiners at patent offices—when examining a patent application—refer to claims in the application document and provides a list of previously published (patent) documents relevant for the application at hand. Furthermore, they often underline passages in these documents to sustain their decisions.

The idea behind organizing this task is to investigate the support an IR system could give an IP expert in retrieving documents and passages relevant to a set of claims. The topics in this task are sets of claims extracted from actual patent application documents. Participants were asked to return documents in the CLEF-IP 2012 corpus and mark the passages relevant to the topic claims.

The participants were provided with a set of 51 training topics. Splitting by the document language of the application document where the topic claims appear, 18 topics were in German, 21 in English and 12 in French. For the test set we have created 105 topics, with 35 in each language. Both training and test topics were created manually. We describe below the steps in obtaining the topics and their relevance judgments.

[2] A list of kind EPO kind codes is listed at https://register.epo.org/espacenet/help?topic=kindcodes.
Kind codes used by the WIPO are listed at http://www.wipo.int/patentscope/en/wo_publication_information/kind_codes.html

**Topic Selection.** Before actually creating the topics we have first created a pool of candidate documents out of which we extracted the sets of topic claims. The documents in the pool were patent applications not occurring in the CLEF-IP data corpus (i.e published after 2001), with content in all main XML fields, and with two to twelve citations in the collection corpus listed in their search report. We have counted only the highly relevant citations (marked with 'X' or 'Y' on the search reports) leaving out the citations about technological background of the invention (marked with 'A' on the search reports). (In the CLEF-IP terminology, we have used thus only highly relevant direct citations occurring in the CLEF-IP corpus [8].) In the CLEF-IP collection, there are few patents with a higher number of highly relevant citations, and these are usually very large documents, with a large patent family, and, by manual inspection, looking for candidate topics, difficult to handle. The next step in creating the topics and

| | | |
|---|---|---|
| | --- | |
| X | WO 01 26573 A (COHERENT INC)<br>19 April 2001 (2001-04-19)<br>* page 13, line 30 - page 15, line 16;<br>figure 3 * | 1-3,7 |
| | --- | |
| Y | EP 1 101 450 A (PULSION MEDICAL SYSTEMS<br>AG) 23 May 2001 (2001-05-23)<br>* page 5, line 9 - line 22; figure 2 * | 8 |
| | --- | |

**Fig. 1.** Extract from a search report.

their qrels consisted of examining the search reports to extract sets of claims and their relevant documents together with the passages in the documents. For each highly relevant citation document in the search report and in our collection, we extracted the claim numbers[3] the citation referred to. These formed the sets of claims for a candidate topic. We then looked at the mentioning of relevant passages in the cited document and decided if the candidate topic could be kept in the set of topics or not. Rejecting a topic candidate was done when:

- relevant documents were referring to figures only;
- no relevant passage specifications were given for the listed citations, or it was mentioned that the whole document is relevant;
- the search report had the mention 'Incomplete search' which usually means that the search done by the patent expert was not done for all claims.

From one patent application document it was possible to extract several sets of claims as topics, often with completely different sets of relevance judgments.

We illustrate how this step was done by an example: Figure 1 shows a part of the search report for the EP-13884446 patent application. The numbers on the right hand side represent claim numbers. Two sets of claims can be identified: {1,2,3,7} and {8}. Leaving the references to figures out, there is enough infor-

[3] Claims in patent documents are numbered for ease of reference

mation to identify relevant passages in the two relevant citations, WO-0126537 and EP-1101450, so we kept the two sets of claims as topics[4].

Concretely, a topic in this task is formulated as:

```
<tid>tPSG-5</tid>
<tfile>EP-1480263-A1.xml</tfile>
<tclaims>/patent-document/claims/claim[1] /patent-document/claims/claim[2]
/patent-document/claims/claim[3] /patent-document/claims/claim[16]
/patent-document/claims/claim[17] /patent-document/claims/claim[18]
</tclaims>
```

The text files with the retrieval results that the participants submitted to this task had the following format:

```
topic_id   Q0   doc_id   rel_psg_xpath   psg_rank   psg_score
```

where:

- `topic_id` is the identifier of a topic
- `Q0` is a value maintained for historical reasons
- `doc_id` is the identifier of the patent document in which the relevant passages occur
- `rel_psg_xpath` is the XPpath identifying the relevant passage in the `doc_id` document
- `psg_rank` is the rank of the passage in the overall list of relevant passages
- `psg_score` is the score of the passage in the (complete) list of relevant passages

Only one XPath per line was allowed in the result files. If more passages are considered relevant for a topic, these have to be placed on separate lines. The maximum number of lines in the result files is limited to containing 100 `doc_ids` when ignoring the XPaths.

**Creating the Relevance Judgments.** Both in the topics and in the relevance judgements, reference to the claims and relevant passages is encoded by means of XPaths. Where the search report referred to specific lines rather than paragraphs, we took as relevant the set of paragraphs fully covering those lines. Once the topics chosen, the last step was to actually create the files with the relevance judgements. We did this manually, by matching the passage indications in the search reports with the content of the patent documents and with the content and XPaths stored in the XML files. To ease this process we have used a system developed in-house, a screenshot of which can be seen in Figure 2. We see in Figure 2 that the qrel generating system has three main areas:

- a topic description area where, after typing in the patent application document identifier (here, EP-1384446-A1), we can assign the topic an identifier (unique in the system), we define the set of claims in the topic, save it, navigate among its relevant documents with the 'Prev' and 'Next' buttons.

[4] In the end, from the EP-13884446 application document we have extracted five topics: PSG30 to PSG34.

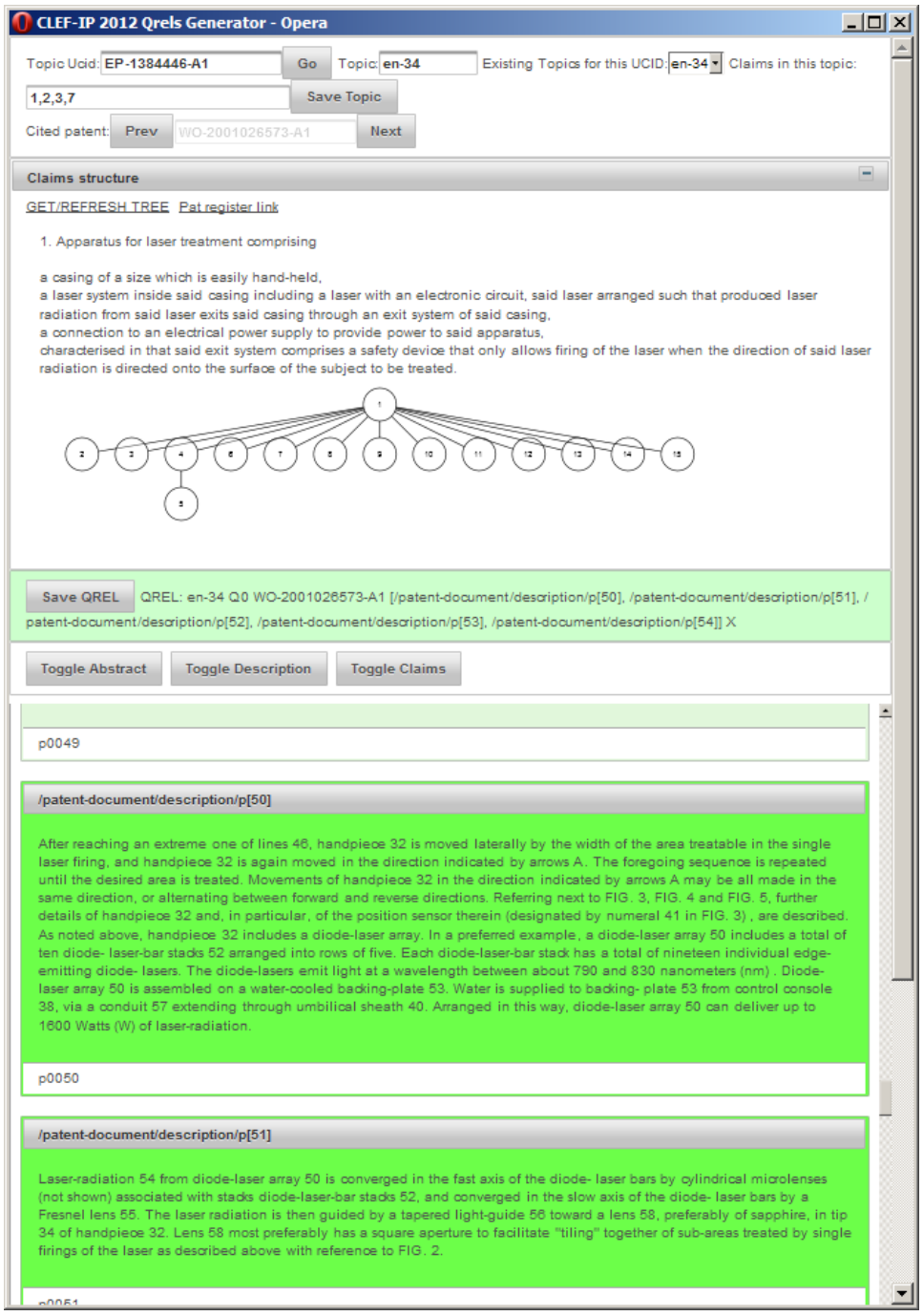


**Fig. 2.** Creating the qrels.

– a claim structure area where we display the claims and the claim tree. Also in this area we give a direct link to the application document on the EPO Patent Register server.
– a qrel definition area where individual passages (corresponding to XPaths in the XML documents) are displayed. Clicking on them will select them to be part of the topic's qrels. For convenience, we provide three buttons by which we can select with one click all of the abstract's, description's or claims' passages. When clicking on the 'Save QREL' button the selected passages are saved in the database as relevant passages for the topic in work.

**Evaluation Measures.** The evaluation of the passage retrieval task was carried out on two levels: those of the document and of the passage. The objective of measuring systems retrieval quality at the document level is to evaluate the system performance in retrieving whole relevant patent document. The objective of the passage level evaluation is to measure the system ranking quality of the relevant passages in the relevant patent documents.

Regarding the document-level evaluation, we focused on recall-oriented retrieval measures. Patent retrieval evaluation score [3] (PRES), Recall, and MAP were used for evaluating the system performance in retrieving the relevant documents to a given topic. The cut-off used for computations was 100 patent documents (not passages!) in the results list.

At the passage-level we measured system performance for ranking the passages in a given relevant document according to their relevance to the topic. The measures of choice are mean average precision MAP and precision. In more detail, the scores are calculated as follows:

For each document relevant to a given topic, we compute the average precision (AP) and precision. To differentiate them from the usual average precision and precision measures calculated at topic level, we call them AP at document level, AP(D), and precision at document level, Precision(D). Then, for a certain topic $T$ and a relevant document $D_i$, the average precision at the $D_i$ document level is computed by the following equation:

$$AP(D_i, T) = \frac{1}{n_p(D_i, T)} \sum Precision(r) \cdot rel(r) \tag{1}$$

where $n_p(D_i, T)$ is the number of relevant passages in the document $D_i$ for the topic $T$, $Precision(r)$ is the precision at rank $r$ in the ranked passage list, and $rel(r)$ is the relevance of the passage, a value in the set {0,1}. The precision at the $D_i$ document level for the topic $T$, $Precision(D_i, T)$, is the percentage of retrieved relevant passages in the list of all retrieved passages for the document $D_i$.

We, thus, compute AP(D) and Precision(D) for all the relevant documents of a given topic, and the average of these scores is calculated to get the score of the given topic:

$$AP(D, T) = \frac{\sum AP(D_i, T)}{n(T)}, \quad Precision(D, T) = \frac{\sum Precision(D_i, T)}{n(T)} \tag{2}$$

where $AP(D,T)$ is the average precision per documents for topic $T$, $Precision(D,T)$ is the precision per documents for topic $T$, $n(T)$ is the number of relevant documents for topic $T$.

Finally, MAP(D) and Precision(D) are computed as the mean of $AP(D,T)$ and $Precision(D,T)$ across all topics.

The MAP(D) and Precision(D) measures carry similarities with the measures used in INEX evaluation track, the 'Relevant in Context' tasks [2] where instead of sequences of characters we look at XPaths.

### 2.2 Flowchart Recognition

Patent images play an important role in the every-day work of IP specialists by helping them make quick decisions on the relevancy of particular documents to a patent document. The design of the Flow Chart Recognition task in the CLEF-IP lab aims at making the content of the patent images searchable and comparable, a topic of high interest for the IP specialists.

The topics in this task are patent images representing flow-charts. Participants in this task were asked to extract the information in these images and return it in a predefined textual format. The set of training topics contained 50 flow-charts together with their textual representation (the qrels). The set of test topics contains 100 flow-charts. All images are black and white tif files. We were not interested, yet, in the connection between the patent images and the patent documents they occur in.

**Topic Selection** Our job in selecting the topics for this task was much easier than in the 'Claims to Passage' task. This is due to the fact that in 2011 the CLEF-IP lab had organized an image classification task, where one of the classification classes was flow-charts. Having already a set of images containing flow-charts, we had to browse through them and select images for the topic sets. Since we chose to model the flow-charts as graphs, we left out from the topic selection images with 'wrong' graph data, like, for example, edges ending in the white.

**Creating Relevance Judgments.** Once the textual representation of the flow-charts was fixed, we have manually created the textual representations for each of the topics in the training and test sets. In Figure 3 we can see an example of a flow-chart image and its textual representation. In the textual representation of a flow-chart, `MT` stands for meta information about the flow-chart (like number of nodes, edges, title), lines starting with `NO` describe the nodes of the graph, lines starting with `DE` describe directed edges, while lines starting with `UE` describe uni-directed edges in the graph. Lines beginning with `CO` denote comments that are not to be automatically processed.

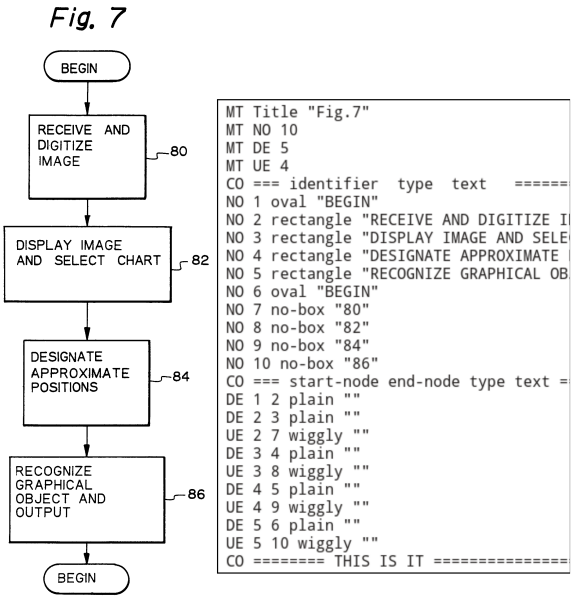


**Fig. 3.** A flow-chart image and its textual representation.

**Evaluation Measure** Since flow-charts can be modeled as graphs, to assess how good the image recognition is done in this specific task, the main evaluation measure is the graph distance metric based on the **mcs**, most common sub-graph (see [1,9]). The distance between the topic flowchart $F_t$ and the submitted flowchart $F_s$ is computed as:

$$d(F_t, F_s) = 1 - \frac{|mcs(F_t, F_s)|}{|F_t| + |F_s| - |mcs(F_t, F_s)|} \tag{3}$$

where $|.|$ denotes the size of the flowchart/graph. In our case, the size of the flowchart is the number of edges plus the number of nodes.

The distance between the topic and the submitted flowcharts is to be computed at three levels:

- basic: only the flowchart structure is taken into consideration, i.e. nodes and edges without type information and without text labels.
- intermediate: use the flowchart structure together with the node types.
- complete: flowchart structure, node types, text labels

To actually compute the distance we use an in-house implementation of the McGregor algorithm for computing most common sub-graphs [4].

In the process of developing these evaluations, we have found that the complete evaluation is better served by a combination of node type matching and edit-distance measuring of the text labels. This is because we cannot make a hard-match between the OCRed text provided by the participants and the one

in the gold standard. Therefore, we allow the McGregor algorithm to generate *all* possible largest common sub-graphs, and compute the best, average and distribution of edit-distances between the nodes of each of these sub-graphs and the gold standard. This is unfortunately extremely time consuming.

### 2.3 Chemical Structure Recognition

Given the importance of chemical molecular diagrams in patents, the ability to extract such diagrams from documents and to recognize them to the extent necessary to automatically compare them for similarity or identity to other diagrams is a potentially very powerful approach to identifying relevant claims. This task was divided into two parts, segmentation and recognition;

**Segmentation** For this sub-task, 30 patents were selected, rendered to 300dpi monochrome multipage TIFF images and all chemical molecular diagrams were manually clipped from the images using a bespoke tool. This clipping tool recorded the minimal bounding box size and coordinates of each diagram clipped and the results recorded in a ground-truth comma separated value (CSV) file. The participants were asked to produce their own results CSV file containing this bounding box clip information for each diagram that their systems could identify.

Another bespoke tool was written to automatically compare the participants' results file with the ground truth file. This identified matches at various tolerance levels, where a match is awarded if every side of a participant's bounding bounding box is within the tolerance number of pixels of the corresponding side of a ground truth bounding box. Evaluation results were calculated for each of a range of tolerances starting at 0 pixels and increasing to the maximum number of pixels that still disallowed any single participant bounding boxes from matching more than one ground truth bounding box. This maximum limit in practice was 55 pixels, or just under 0.5 cm.

The number of true positive, false positive and false negative matches were counted for each tolerance setting, and from that the precision, recall and $F_1$-measure was calculated.

**Recognition** A diagram recognition task requires the participants to take a set of diagrams, analyse them to some recognised format and submit their recognised format files for evaluation. In order to evaluate the results, these submitted format files must be compared to a ground-truth set of format files. Therein lies a difficult problem with respect to the chemical diagram recognition task for patent documents. The currently most complete standard format for chemical diagrams is the MOL file format. This format captures quite well fully specified chemical diagram molecules. However, it has become standard in patent documents to describe whole families or classes of molcules using diagrams that extend standard molecule diagrams with graphical representations of varying structures, called *Markush structures*. Markush structures cannot be represented in MOL files.

Given the standard nature of MOL files, there have been a significant number of research and commercial projects to recognise diagrams with all the features that can be represented by MOL files. However, without standard extensions to MOL files to cope with Markush stuctures, there has been relatively little effort expended in recognising such extended diagrams. With the intention of fostering such efforts, the chemical structure recognition task for CLEF-IP 2012 was designed to expose participants to a relatively small number of the simpler of these extended structures, while also providing a large number of cases fully covered by the current MOL file standard.

A total of 865 diagram images, called the *automatic set*, were selected. The diagrams in this set were fully representable in standard MOL files. Evaluation of this set was carried out by automatically comparing the participants submitted MOL files with the ground truth MOL files using the open source chemistry toolbox, OpenBabel. The key tool in this set is the InChi representation (International Chemical Identifier). OpenBabel was chosen among other tools offering similar functionality because it is free and available to everyone. The number of correctly matched diagrams (and the percentage correctly matched) were reported for each participant.

A *manual set* of 95 images were chosen which contain some amount of variability in their structure and which can only be represented in MOL files by some abuse of the MOL file standard. These can not be automatically evaluated as the OpenBabel system cannot deal with the resulting structures. However, such MOL files can still be rendered to an image format using the MarvinView tool from ChemAxon. Thus it was possibile to carry out the evaluation of this set by manual visual comparison of the original image, the MarvinView generated image of the ground-truth MOL file for the image, and the MarvinView generated image of the participant's submitted MOL file. To this end a bespoke web application was written to enable the organisers and participants to verify the manual visual evaluation.

It was less than satisfactory to have to carry out the evaluation of this latter set manually, and even more so that we had to exclude from the set structures that appear in patent files but which cannot be rendered from (abused) MOL files using MarvinView. This points strongly to a need in the community to develop either an extension or an alternative to MOL files that can fully support common Markush structures together with the necessary ancillary tools for manipulating, comparing and rendering such structures to images.

## 3 Submissions and Results

Table 1 gives a list of institutions that submitted experiments to the CLEF-IP lab in 2012. Research collaborations between institutions can be identified by the RunID. Note that two different groups from the Vienna University of Technology have each contributed to differently identified submissions (`tuw` and `lut`).

**Table 1.** List of participants and runs submitted

| RunID | Institution | | CLM | FC | CS |
|---|---|---|---|---|---|
| uob | University of Birmingham, School of Computer Science | UK | | | x |
| bit | Univ. of Applied Sciences, Information Studies, Geneva | CH | x | | |
| chm | Chemnitz University of Technology, Department of Computer Science | DE | x | | |
| cvc | Computer Vision Center, Universitat Autónoma de Barcelona | ES | | x | |
| hild | Univ. Hildesheim, Information Science | DE | x | | |
| humb-inr | Humbold Univ., Dept, of German Language and Linguistics | DE | | x | |
| humb-inr | INRIA | FR | | x | |
| joann | Joanneum Research, Institute for Information and Communication Technologies | AT | | x | |
| lut | University of Lugano | CH | x | | |
| tuw | Univ. of Macedonia, Department of Applied Informatics, Thessaloniki | GR | x | | |
| saic | Chemical Biology Laboratory, SAIC-Frederick Inc. | US | | | x |
| lut | Vienna University of Technology, Inst. for Software Technology and Interactive Systems | AT | x | | |
| tuw | Vienna University of Technology, Inst. for Software Technology and Interactive Systems | AT | x | | |
| tuw | Univ. of Wolverhampton, School of Technology | UK | x | | |
| | **Total:** | | 31 | 13 | 7 |

### 3.1 Evaluation Results

**Claims to Passage** As stated in section 2.1, we computed two sets of measures, one at the document level, very similar to the measurements done in the last years, and one at the passage level. To compute measures at the document level we have ignored the passage information in the participants' submissions and kept only the <topic, relevant document, rank> tuples. On these we have computed PRES, Recall and MAP, both for the complete set of topics, as well as split on languages. At the passage level, we have computed MAP- and Precision-like measures, by computing passage AP, respectively passage Precision, for each relevant document, then averaging over the topic's relevant documents. The final scores are obtained by averaging over all queries.

The solutions chosen by the submitting participants range from two-step retrieval approaches, namely a document level retrieval in the first step and a passage level retrieval in the second step (*bit* and *chm*) to using Natural Language Processing techniques (*lut*). The *tuw* team used a distributed IR system by splitting the CLEF-IP collection by IPC codes, while the *hild* team experimented with search types trigram-based searches. All participants have used translation tools on the generated queries.

Figure 4 presents the PRES, MAP and Recall at the document level, Figure 5 shows the Precision and MAP at the passage levels for the complete set of topics. The *tuw* participant was left out in the passage level evaluations because they have submitted experiments referring to documents only, and not passages.

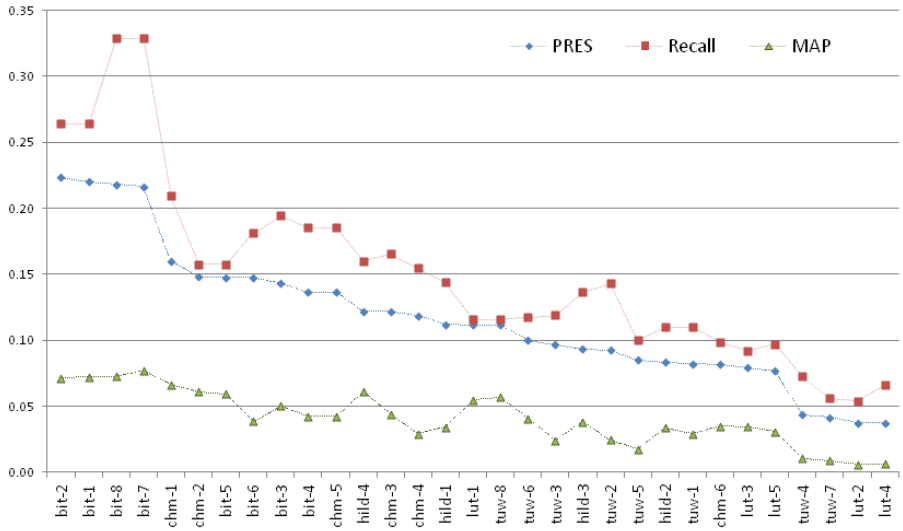


**Fig. 4.** Measures at relevant document level.

Precision(D) and MAP(D) gives an indication about the system performance in ranking the passages in the relevant documents regardless to the quality of the document-level retrieval quality.

**Flow Chart Recognition** Unfortunately, at the time of writing these workshop notes, the execution of the evaluation programm was not closed. We will post the results on the project's website, `http://www.ifs.tuwien.ac.at/~clef-ip`

**Chemical Structure Recognition** Only one participant, *saic*, submitted an attempt at the chemical molecular diagram segmentation subtask. They submitted two attempts, both using as input the multi-page tiff files provided by the organisers. The difference was that in one run they used `tiffsplit` to separate the pages into individual files, while in the other one they used OSRA native file reading/rendering capability. They achieved significantly better performance on the latter with results presented in Table 2 (note that a tolerance of 55 is just under 0.5cm): Both *saic* and *uob* submitted result sets (1 and 4 respectively) for the diagram recognition sub-task (Table 3).

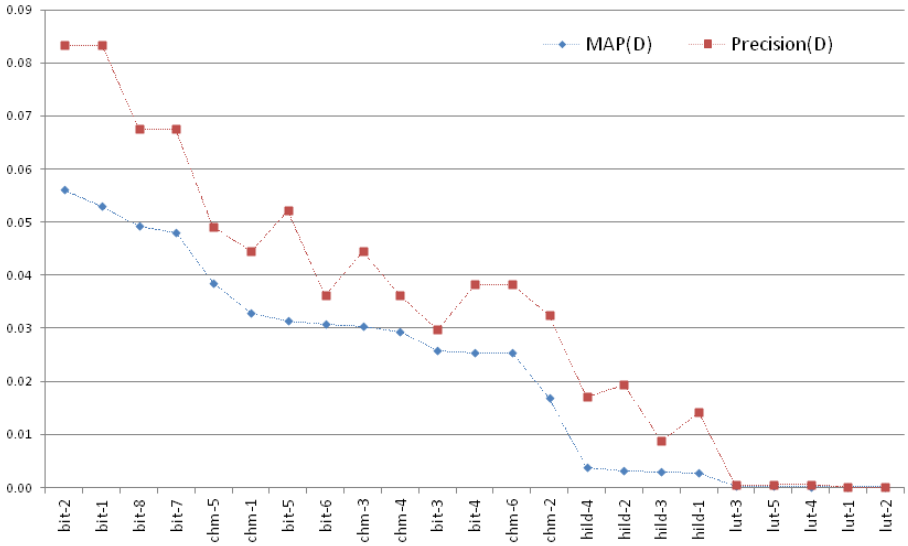


**Fig. 5.** Measures at relevant passage level.

**Table 2.** Chemical molecular diagram segmentation results.

| Tolerance | Precision | Recall | $F_1$ |
|---|---|---|---|
| 0 | 0.70803 | 0.68622 | 0.69696 |
| 10 | 0.79311 | 0.76868 | 0.78070 |
| 20 | 0.82071 | 0.79543 | 0.80787 |
| 40 | 0.86696 | 0.84025 | 0.85340 |
| 55 | 0.88694 | 0.85962 | 0.87307 |

**Table 3.** Chemical diagram recognition results.

| | **Automatic Set** | | | **Manual Set** | | | **Total** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **#Structures** | **Recalled** | **%** | **#Structures** | **Recalled** | **%** | **#Structures** | **Recalled** | **%** |
| saic | 865 | 761 | 88% | 95 | 38 | 40% | 960 | 799 | 83% |
| uob-1 | 865 | 832 | 96% | 95 | 44 | 46% | 960 | 876 | 91% |
| uob-2 | 865 | 821 | 95% | 95 | 56 | 59% | 960 | 877 | 91% |
| uob-3 | 865 | 821 | 95% | 95 | 44 | 46% | 960 | 865 | 90% |
| uob-4 | 865 | 832 | 96% | 95 | 54 | 57% | 960 | 886 | 92% |

Clearly, both groups, unsurprisingly, found the diagrams with varying elements significantly more challenging than the more standard fixed diagrams.

## 4 Final Observations

We have described the benchmarking activities done in the frame of the CLEF-IP 2012. One of the main challenges faced by the organizers were obtaining relevance judgments and choosing topics for the 'Passage Retrieval Starting from Claims' task. The effort spent on this challenge prompted us to waive an additional pilot task originally proposed for this year, in which we were interested in finding description passages relevant to a claim in a patent application document.

Another challenge was finding proper measures to assess the efficiency of the passage retrieval task as formulated in the CLEF-IP lab and for the 'Flow-chart Recognition' task. The proposed measures are to be a starting point for further discussions on what is the best way to assess the effectiveness of these types of information retrieval.

**Acknowledgments** This work was partly supported by the EU Network of Excellence PROMISE (FP7-258191) and the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx[5] (No. 825846).

## References

1. Horst Bunke and Kim Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, 1998.
2. Jaap Kamps, Jovan Pehcevski, Gabriella Kazai, Mounia Lalmas, and Stephen Robertson. Inex 2007 evaluation measures. In *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers*, volume 4862 of *Lecture Notes in Computer Science*, pages 24–33. Springer, 2008.
3. W. Magdy and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010*, 2010.
4. James J. McGregor. Backtrack search algorithms and the maximal common subgraph problem. *Softw., Pract. Exper.*, 12(1):23–34, 1982.
5. F. Piroi, M. Lupu, and A. Hanbury. Effects of Language and Topic Size in Patent IR: An Empirical Study. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics, Third International Conference of the CLEF Initiative, CLEF 2012, Rome, Italy, September 17-20, 2012, Proceedings*, volume 7488 of *Lecture Notes in Computer Science*. Springer, September 2012.
6. F. Piroi, M Lupu, A. Hanbury, and V. Zenz. CLEF-IP 2011: Retrieval in the intellectual property domain, September 2011.
7. F. Piroi and J. Tait. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. Technical Report IRF–TR–2010–00005, Information Retrieval Facility, Vienna, September 2010. Also available as a Notebook Paper of the CLEF 2010 Informal Proceedings.

[5] http://www.joanneum.at/?id=3922

8. G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In C. Peters, G.M. Di Nunzio, M. Kurimo, D. Mostefa, A. Penas, and G. Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009*, volume 6241, pages 385–409. Springer, 2010.
9. Walter D. Wallis, Peter Shoubridge, Miro Kraetzl, and D. Ray. Graph distances using graph union. *Pattern Recognition Letters*, 22(6/7):701–704, 2001.