# CLEF–IP 2010
# Track Guidelines

Florina Piroi

Information Retrieval Facility
Vienna

## 1 Introduction

The CLEF–IP track was launched in 2009 to investigate IR techniques for patent retrieval and it was part of the CLEF 2009 evaluation campaign. In 2010, the track continues as a benchmarking activity of the CLEF 2010 conference.

The track utilizes a collection of more than 1.3 million patent documents derived from EPO sources. The collection covers English, French and German with at least 150,000 documents in each language.

There are two tasks in the 2010's track. The first one is to find patent documents that are candidates to constitute prior art for a given document. The second task is to classify a given document according to the International Patent Classification system (IPC)[1]. As in the track's first year, relevance judgements will be produced using the patent citations.

## 2 Target data

The target data set contains all EPO documents that have an application date previous to 2002. (2,680,698 patent documents constituting 1,331,106 patents, with a disk size of 84GB). The same as in 2009, the Test Collection Corpus was delivered to the participants "as is", without merging the documents related to the same patent into one document. Each patent is identified by a unique patent number– a string starting with "EP" and followed by 7 digits. Corresponding to

[1] See http://www.wipo.int/classifications/ipc/en/

each patent is a directory containing the patent documents related to that patent. The layout is nnnnnn/nn/nn/nn/*.xml.

For example, to patent EP 0981201 corresponds the directory containing files EP-0981201-A2.xml, EP-0981201-A3.xml, and EP-0981201-B1.xml:

```
> pwd
/000000/98/12/01
> ls
EP-0981201-A2.xml EP-0981201-A3.xml EP-0981201-B1.xml
```

### Patent documents

In general, one patent (identified by a unique patent number) corresponds to several patent documents generated at different stages of the patent's life-cycle. Each stage is denoted by a kind code together with a version number[2]. The most common kind codes[3] are the following:

| | |
|---|---|
| A1 | publication of application with search report |
| A2 | publication of application without search report |
| A3 | publication of search report |
| A4 | supplementary search report |
| A8 | corrected title page of an EP–A document |
| A9 | complete reprint of an EP–A document |
| B1 | granted patent |
| B2 | granted patent after modification |
| B8 | corrected front page of an EP–B document |
| B9 | complete reprint of an EP–B document |

The "B" kind codes mark granted patents, and they necessarily have been earlier published as an "A1" or "A2" document.

[2] For the EP patents, documents at different stages have the same numeric identifier. For other patent offices this is not always the case. For example, the patent document US-6689545-B2 represents a US granted patent with its application document publication number US-2003011722-A1

[3] A full list of kind codes can be found at http://tinyurl.com/EPO-kindcodes

## 3 The Prior Art Candidate Search Task

The first task of the CLEF–IP track (PAC) consists in finding patent documents in the target collection that may invalidate a given patent application. In intellectual property language: find documents that may constitute prior art for a topic patent.[4] The topic document is a patent application document, A1 or A2, where the citation information was removed.

Differently from last year's topics, where we have created a virtual patent document with claims in German, English and French, we have not taken extra care that claims are present in all three languages. This year's topic documents usually contain the claims in only one of the three languages, with about 67% of documents having Enligsh content, 26% German content, and 7% French content.

A PAC topic file contains a concatenation of topics. The structure of one topic is as follows:

```
<topic>
<num> PAC-number</num>
<narr> Find all patents in the collection that potentially
invalidate patent application patentNumber. </narr>
<file> fileName.xml </file>
</topic>
```

where *number* is the number of the topic (between 1 and 300 for the training set), *patentNumber* is "EPÂťÂť followed by the seven digits and a patent kind code (A1 or A2), and *fileName.xml* is a concatenation of `PAC-`*number* and *patentNumber* with an xml file extension. *fileName.xml* can be found in the same folder where the topic file is.

There are two PAC topics sets, one with 2,000 topics and one with 500 topics which is a subset of the first one. We draw attention that the subset of 500 topics is not just a listing of the first 500 entries in the larger topic set. The `<num>` field of the small set of PAC topics contains `PACs-`*number*, to differentiate it from the topics in the larger set. Participants are asked to submit runs at least for the small topic set.

[4] For an explanation of what prior art is, please refer to the discussion on the track's web page.

## 4 The Classification Task

This task is new in the CLEF–IP evaluation track and its requirement is to classify a given topic document according to the International Patent Classification system (IPC)[5], up to the SUBCLASS level. The classification (CLS) topic file contains a concatenation of topics. The structure of one CLS topic is as follows:

```
<topic>
<num> CLS-number</num>
<narr> Classify patent document patentNumber
according to the IPC system.</narr>
<file> fileName.xml </file>
</topic>
```

where *number* is the number of the topic, *patentNumber* is "EP" followed by the seven digits and a patent kind code, and *fileName.xml* is a concatenation of `CLS-`*number* and *patentNumber* with an xml file extension. *fileName.xml* can be found in the same folder where the topic file is.

There is one set of classification topics, containing 2,000 topics.

## 5 The Training Set

Similar to last year, we have released a set of topics, together with their relevances for training purposes. The participants can use this set to train and tune their systems. The CLEF–IP 2010 training set contains documents and relevance assessments for 300 topics similar to the PAC search task. The relevance assessments are automatically extracted from patent citation information, and have the following format:

```
PACt-101    EP-1179726-A2    2
PACt-102    EP-0916655-A2    1
PACt-102    EP-0927717-A1    1
```

where the first column contains the topic number, with `PACt` marking that it is a training topic, the second column contains the relevant

[5] For a quick and at the same time rich introduction to the IPC system see http://www.wipo.int/classifications/ipc/en/faq/

document as identified by the document's file name without the extendion, and the last column is the relevance degree of the document (2 being more relevant than 1).

There is no training set for the classification task, as participants should use the target data to train their classifier.