CrossMark

Analysis of the patent documentation coverage of the CPC in comparison with the IPC with a focus on Asian documentation

Bart Degroote[*], Pierre Held

European Patent Office, Patentlaan 2, 2288EE Rijswijk, The Netherlands

ARTICLE INFO

ABSTRACT

In this study we analyse the patent documentation coverage of the Cooperative Patent Classification system (CPC) in comparison with the coverage of the International Patent Classification system (IPC). We demonstrate that for patent searchers the coverage in terms of patent families is more relevant than the coverage in terms of patent documents. Of all the patent families classified in the IPC only about one third thereof is currently classified in the CPC. Most of the patent families that are not classified in the CPC are patent families including published patent applications filed at the three major Asian patent offices (JPO, KIPO and SIPO). We demonstrate a method of comparing the CPC patent documentation coverage to the IPC patent documentation coverage down to IPC subgroup level and how this can be applied to improve the efficiency of classification based search. The coverage of the CPC relative to the IPC must be understood in terms of: the CPC being assigned to families, and the IPC being assigned to individual documents. Furthermore, the CPC has only been implemented relatively recently compared to the IPC, with SIPO and KIPO focusing on the classification of their front files in the CPC.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Patent classification systems are widely used by patent examiners and patent information experts for retrieving relevant prior art. The most commonly used classification systems are the International Patent Classification system (IPC), the Cooperative Patent Classification (CPC) and the FI (File Index) and F-term (File forming term) classification. When using a certain classification system for search, one needs to be aware of the documentation coverage of that classification system. For example, a patent searcher may be searching for prior art using the relevant CPC symbol(s) but may miss a relevant document because this document is not classified in the CPC.

Most of the patent publications and utility models filed in a WIPO member are classified in the IPC. For this reason, the IPC has a large patent documentation coverage. On the other hand, differences in classification practise between various patent offices can make document retrieval based on the IPC more complicated. For example, a patent searcher may be searching for prior art using an IPC symbol that is considered relevant by the European Patent Office (EPO) but he may miss a relevant document because this document is a Japanese patent application that was classified in the IPC by the Japanese Patent Office (JPO) with a different symbol.

The CPC is an extension of the IPC and is jointly managed by the EPO and the US Patent and Trademark Office (USPTO). The CPC has the advantage that it is much more detailed than the IPC. For example, H05K3/34 has no subgroups in the IPC whereas it has many subgroups in the CPC. On the other hand, not every WIPO member state is classifying in the CPC, which means that the documentation coverage is less than for the IPC. Since the CPC came into force on the 1st of January 2013, a growing number of national patent offices have decided to classify their applications in the CPC, including the State Intellectual Property Office of China (SIPO) and the Korean Intellectual Property Office (KIPO). With more offices joining the CPC, the documentation coverage is expected to improve. However, in the current situation it remains important to be well aware of the extent to which the CPC is covering patent documentation in comparison with the IPC.

2. Data collection

The data for this study have been collected from "EPODOC", the EPO's bibliographic database containing about 70M(illion) records classified in the IPC. The database content is the same as "DOCDB",

---

* Corresponding author.
   E-mail addresses: bdegroote@epo.org (B. Degroote), pheld@epo.org (P. Held).

which is the database that is used as the data source for Espacenet, the free online search service from the EPO. The numerical results in this study were obtained in March 2016.

3. Results and discussion

3.1. CPC coverage in terms of patent documents

In Fig. 1 the number of patent documents classified in the CPC is shown per technical field (dark grey). On top thereof, the number of patent documents classified in the IPC in the same technical field, but not being classified in the CPC is shown (light grey). For example, a patent searcher that is searching in the field of Electricity ("Section" H) with CPC symbols would cover about 9.2M patent documents. However, as he limits himself to CPC classified patent documents, he would disregard 6.1M patent documents that are classified in Section H via an IPC symbol. Based on Fig. 1, one would conclude that the CPC patent documentation coverage is best in the field of chemistry and metallurgy (Section C), whereas it is the worst in the field of fixed constructions (Section E).

The overall picture is that out of the 70M patent documents classified in the IPC, there are about 40M classified in the CPC. Out of the 30M patent documents that are not classified in the CPC (only in the IPC), 24M patent documents are patent publications from national applications filed at the JPO, KIPO and SIPO. This means that in terms of patent documents, the CPC covers somewhat less than 2/3 of the patent documents that the IPC covers.

It is noted that in Fig. 1 the sum of the number of patent documents classified in the CPC for each field (total of the dark grey part) is more than 30M because many patent documents are classified in more than one field. The same applies to the patent documents classified in the IPC in a certain technical field and not being classified in the CPC (total of the light grey part).

3.2. CPC coverage in terms of patent families

At the EPO and USPTO, the CPC is applied at simple patent family level, i.e. all members of the same simple patent family (i.e. sharing exactly the same set of active priorities) share the same CPC classification picture. A patent family therefore corresponds to a group of patent documents that cover the same (or similar) information. Frontfile IPC symbol allocation is done at document level. This means that a patent family may contain documents with different IPC symbols allocated. In such a case we consider for the purpose of this article that the IPC classification of the patent family corresponds to all the IPC symbols given for the different documents. It is noted that in this study when "the number of patent families" is being referred to, this number has been obtained by an analysis at the simple patent family level.

Under the assumption that one can search both at family and document level (most search engines do not foresee a search of the IPC at family level), it may be more relevant for a patent searcher to know how many patent families are covered in the CPC, rather than how many patent documents are covered in the CPC. As all patent documents in a patent family correspond to the same (or similar) information, the number of patent families covered in a search is likely to be a better indicator for the different pieces of information that have been considered in the search.

In Fig. 2 the number of patent families classified in the CPC is shown per technical field (dark grey), together with the number of patent families classified in the IPC in the same technical field, but not classified in the CPC (light grey). In terms of coverage the situation for patent families is quite different from that of patent documents (see Fig. 1). For example, a patent searcher that is searching in Section H with CPC symbols would cover about 3.2M patent families. However, as he limits himself to CPC classified patent families, he would disregard 5.9M patent families that are classified in Section H via an IPC symbol. In other words, based on the patent family analysis, it appears that the CPC covers less of the patent documentation than what one would expect from an analysis based on patent documents.

The difference between an analysis in terms of patent documents and patent families is caused by the fact that many of the patent families classified in the CPC have several "family members" (on average between 2 and 3.5 depending on the technical field) whereas the patent families that are not classified in the CPC and
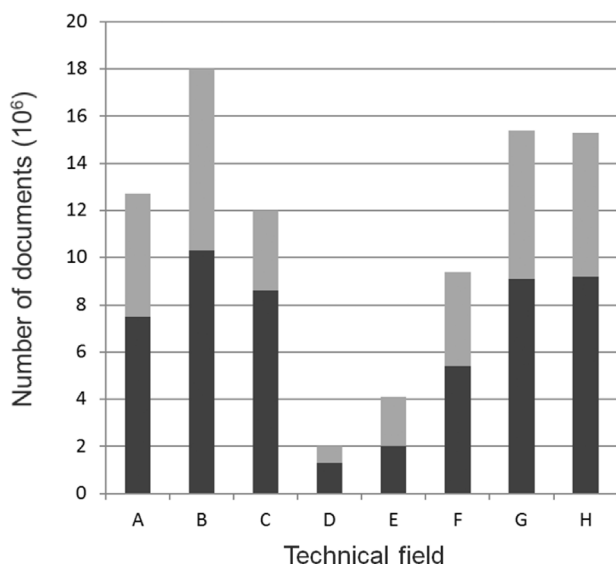

Fig. 1. The number of patent documents classified in the CPC (dark grey) per technical field compared to the number of patent documents classified in the IPC in that field but not classified in the CPC (light grey) (Sections: A - human necessities; B - performing operations, transporting; C - chemistry, metallurgy; D - textiles, paper; E - fixed constructions; F - mechanical engineering, lighting, heating, weapons, blasting; G - physics; H - electricity).
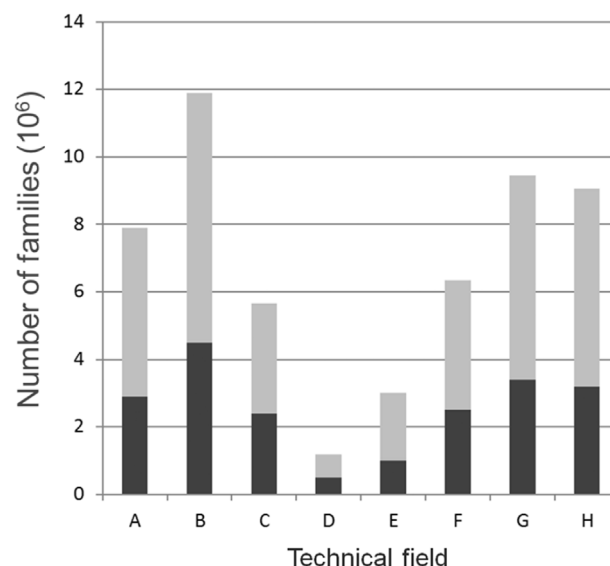

Fig. 2. The number of patent families classified in the CPC (dark grey) per technical field compared to the number of patent families classified in the IPC in that field but not classified in the CPC (light grey).

only classified in the IPC on average only contain one patent document. The overall picture is that out of 45M patent families classified in the IPC, there are about 16M classified in the CPC. Out of the 29M patent families that are not classified in the CPC, 23M patent documents are patent publications from national applications filed at the JPO, KIPO and SIPO. This means that in terms of patent families, the CPC covers somewhat more than 1/3 of the patent documentation that the IPC covers (see also Table 1). This is nevertheless encouraging since the CPC has only been implemented for a relatively short time (since 1.1.2013) with SIPO and KIPO focusing on the classification of their front file document as their first priority.

3.3. CPC coverage of Asian documentation

As mentioned before, a large majority of the patent documents that are not classified in the CPC are patent publications from national applications filed at the JPO, KIPO and SIPO. In Fig. 3 the number of published patent applications (including utility models) filed at the JPO, KIPO and SIPO (across all technical fields) are shown. About 20% of the 17M published Japanese patent applications filed at the JPO via the national route are classified in the CPC. Most of these applications are part of patent families that contain a CPC classified family member (e.g. a US, EP, DE, FR or GB patent application). Some of these were classified by the EPO into ECLA in the past, then the ECLA allocations were converted into CPC allocations. The number of KR patent publications (3.1M) is considerably smaller in absolute numbers than JP publications but a relative large fraction thereof (about 35%) is classified in the CPC indirectly via their patent family. The number of CN patent publications (10M) is somewhat smaller in absolute numbers than JP publications and about 16% thereof are classified in the CPC. Fig. 3 also shows the number of WO publications filed at JPO, KIPO and SIPO: they are relatively small in numbers and all of them are classified in the CPC.

KIPO and SIPO recently joined the CPC, meaning that these offices are now classifying their patent applications in the CPC (in all fields since January 2015 for KIPO and January 2016 for SIPO). As more of these patent applications are published, the fraction of publications from KIPO and SIPO that is classified in the CPC will gradually increase. Below the yearly planning projection of the Korean national documents with CPC allocations classified by KIPO (Table 2) as well as the yearly planning projection of Chinese national documents with CPC allocations classified by SIPO (Table 3) is shown [1].

As can be seen from Table 2, all "Front File" applications (i.e. newly filed and not yet published patent applications) filed at KIPO are classified in the CPC from 2015 onwards. KIPO will also classify a large proportion of the "Backfile" documents (i.e. already published patent applications), meaning that all the 2M KR documents shown in Fig. 3 that are not classified in the CPC yet will be classified in the CPC in the near future. Currently 0.22M Korean national patent publications that are classified by KIPO are available in EPODOC. It is noted that Table 2 refers to patent applications and utility models.

As can be seen from Table 3, all front file applications filed at SIPO are being classified in the CPC from 2016 onwards. Only a
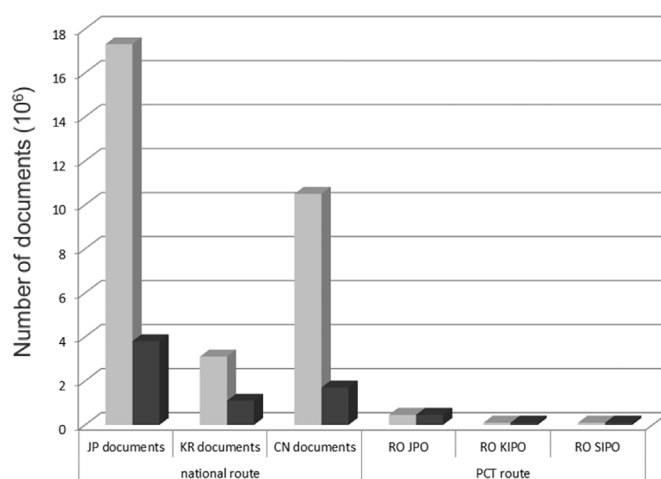

Fig. 3. Number of published patent applications filed at JPO, KIPO and SIPO classified in the IPC (light grey) and the CPC (dark grey) filed via the national route and the PCT route.

Table 2
The yearly planning projection of the Korean national documents with CPC allocations classified by KIPO [1].

Year	Back file Documents to be Classified in CPC	Front File Applications to be Classified in CPC
2015	232,000	All
2016	457,000	All
2017	390,000	All
2018	390,000	All
From 2019	N/A	ALL

limited number of backfile documents have been classified so far and it depends on the availability of resources to what extent more backfile documents will be classified in the future. SIPO is considering the reclassification of its backfile, however with more than one million applications received per year that need to be classified in the CPC, it is understandable that it may take time for SIPO to classify its full backfile in the CPC. As a consequence, until further notice, a large portion of the 8.8M CN documents shown in Fig. 3, which are currently not classified in the CPC, may not be classified in the CPC in the near future. It is noted that Table 3 only refers to patent applications, not utility models. Subject to availability of resources, SIPO will strive to also classify utility models in the CPC in the future. Currently 0.25M Chinese national patent publications that are classified by SIPO are available in EPODOC.

Based on these results one can conclude that the next step in improving the CPC coverage would be the inclusion therein of JP patent documents, as otherwise it will remain restricted to families containing (a) CPC-classified family member(s). Efforts are made within the IP5 cooperation framework aiming at aligning the CPC, FI/F-terms and the IPC. Furthermore, the JPO is adopting the CPC in a growing number of technical areas. In these cases, although the FI/F-term system needs to be searched, the corresponding classification entries are the same as the ones under the CPC. The same

Table 1
Overview of the number of patent documents and patent families classified in the IPC and the CPC.

EPODOC	Number of patent documents (Millions)	Number of patent families (calculated -Millions)
Classified in the IPC	70	45
Classified in the CPC and the IPC	40	16
Not classified in the CPC, only classified in the IPC	30	29

Table 3
The yearly planning projection of the Chinese national documents with CPC allocations classified by SIPO [1].

	Year	Back file Documents to be Classified in CPC	Front File Applications to be Classified in CPC
Status	2014	89,000	155,820
	2015	380,000	1100,000
plan	2016	Deliver CPC data on a monthly basis until all the CPC data produced by SIPO in 2015 have been sent.	All
	2017	Depending on the availability of resources	

happens following IPC changes, where the FI system is adapted accordingly. The current CPC coverage of KR patent documents is about 35% and will increase in the near future due to the classification into the CPC of all new applications from 2015 onwards and of the whole backfile by the end of 2019. The current CPC coverage of CN patent documents is of about 16%. Although all new applications from 2016 onwards are being classified in the CPC by SIPO, a large portion of the backfile may remain unclassified in the CPC for a number of years.

3.4. Case study: field dependent analysis of the CPC coverage

The results shown above only give the general picture of the CPC coverage. However, when searching for some particular prior art, a patent searcher usually tries to limit himself to documents that are classified in a very particular field, i.e. down to the "main group" level (A61K9/00) or a "subgroup" thereof (A61K9/10) rather than searching at "Section" level (A). The CPC coverage can drastically vary at Main or Subgroup level compared to the coverage at section level. For this reason, patent searchers are encouraged to make their own assessment of the CPC coverage whenever they start searching in a field they are not familiar with. In this case study, we will look at the CPC coverage in the field covered by subgroup H01L 33/56. This group relates to materials for encapsulation of semiconductor dies in Light Emitting Diode (LED) packages. To analyse the CPC and IPC coverage for this subgroup we have used Espacenet. However, the same approach can be done with any other data source, as long as it contains bibliographic patent data as well as classification data for the CPC and the IPC.

In order to collect the numeric data from Espacenet, we used the following search queries:

The first query provides the number of documents classified in the CPC with the subgroup symbol H01L 33/56.

The second query provides the number of documents classified in the IPC with the subgroup symbol H01L 33/56 and not classified in the CPC (meaning not classified in any of sections A-H).

The third to fifth queries provide the number JP, KR and CN documents obtained for the second query.

In Figs. 4 and 5 screenshots from Espacenet are shown for entering the queries.

In Table 4 the results are shown for the queries mentioned above. As mentioned before, preferably the analysis should be done based on the number of families. However, Espacenet does not have the functionality to convert a list of documents into a list of families. For this reason, we have applied an average conversion factor: in Section H a patent family that is classified in the CPC contains on average 2.85 patent documents whereas a patent family that is only classified in the IPC contains on average 1.04 patent documents (see Figs. 1 and 2).

In Fig. 6 the estimated CPC coverage for H01L 33/56 is shown using the above-mentioned conversion factors: about 37% of the patent families is classified in H01L 33/56 in the CPC whereas 63% is classified in H01L 33/56 in the IPC and is not classified in the CPC. About 70% of the patent families that are not classified in the CPC are JP documents. This means that searching with the FI and F-term classification will cover a large part of the documentation that is not classified in the CPC. About 25% of the patent families that are not classified in the CPC are CN documents. Only a minor part of 3% of the patent families that are not classified in the CPC are KR documents and another 3% are patent documents from other national offices.

```
(1)  cpc=H01L33/56

(2)  ipc =H01L33/56 and not (((((((cpc = A or cpc = B)
or cpc = C) or cpc = D) or cpc = E) or cpc = F) or
cpc = G) or cpc = H)

(3)  ipc =H01L33/56 and pn=jp not (((((((cpc = A or
cpc = B) or cpc = C) or cpc = D) or cpc = E) or cpc
= F) or cpc = G) or cpc = H)

(4)  ipc =H01L33/56 and pn=kr not (((((((cpc = A or
cpc = B) or cpc = C) or cpc = D) or cpc = E) or cpc
= F) or cpc = G) or cpc = H)

(5)  ipc =H01L33/56 and pn=cn not (((((((cpc = A or
cpc = B) or cpc = C) or cpc = D) or cpc = E) or cpc
= F) or cpc = G) or cpc = H)
```

Fig. 4. Screenshot from Espacenet for entering the query that provides the number of documents classified in the CPC with the Subgroup symbol H01L33/56.

Fig. 5. Screenshot from Espacenet with the results obtained for the query shown in Fig. 4. The important passage is "Approximately 6235" which corresponds to the number of documents classified in the CPC with the subgroup symbol H01L 33/56.

Table 4
Results from the queries obtained via Espacenet.

Query	Number of documents
cpc = H01L33/56	6235
ipc = H01L33/56 not (((((((cpc = A or cpc = B) or cpc = C) or cpc = D) or cpc = E) or cpc = F) or cpc = G) or cpc = H)	3819
ipc =H01L33/56 and pn=jp not (((((((cpc = A or cpc = B) or cpc = C) or cpc = D) or cpc = E) or cpc = F) or cpc = G) or cpc = H)	2658
ipc =H01L33/56 and pn=cn not (...)	956
ipc =H01L33/56 and pn=kr not (...)	107

3.5. Additional remarks

When analysing the CPC coverage, there are some important issues to be aware of:

1. The CPC has only been fully operational for a few years compared with decades for the IPC and is aimed at advanced users
2. The analysis only works up to the lowest level of classification for the IPC. As mentioned before, the CPC is much more detailed than the IPC. For example, H05K 3/34 has no subgroups in the IPC whereas it has many subgroups in the CPC. As a consequence, the CPC coverage analysis can only be done down to the level of H05K 3/34. Furthermore, to obtain the number of

documents/families classified in the CPC, one should count all the documents/families in H05K 3/34 and in the subgroups thereof.

3. Espacenet does currently not allow analysing large numbers of documents. When a query results in more than 10.000 documents, one obtains the message: "More than 10,000 results found". This problem can be avoided by using another data source, e.g. Questel Orbit, Thomson Reuters, STN/CAS, etc.
4. A symbol can be allocated as "invention information" or "additional information". Invention information is the information in the classified document that is considered as contributing over the state of the art. Additional information is non-trivial technical information which does not in itself represent an addition to the state of the art but might constitute useful information for the searcher. There is also the concept of "Unreviewed" and

patent families in H01L33/56

■ classified in CPC

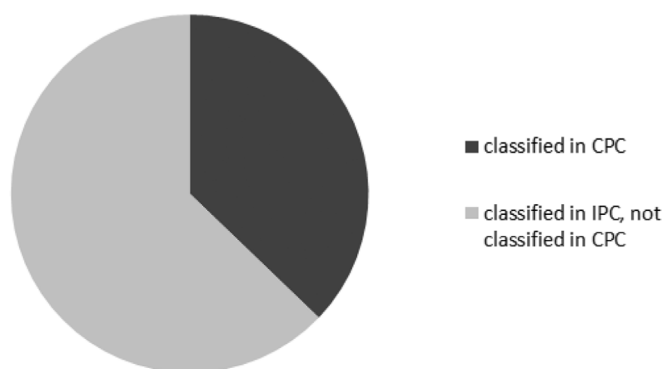■ classified in IPC, not classified in CPC

Fig. 6. Estimate of the CPC coverage for H01L 33/56 based on numerical analysis from Espacenet (using conversion factors of 2.85 for CPC classified patent families and 1.04 for patent families not classified in the CPC).

"Raise-Hand" in the CPC at the EPO. When a symbol is allocated, i.e. proposed, by an unauthorised classifier or the USPTO and the symbol is intended to be reviewed by an authorised classifier, the symbol is considered "Unreviewed" [2]. Furthermore, the EPO (resp. USPTO) cannot delete an allocation provided by the USPTO (resp. EPO): a "Raise-Hand Flag" (RHF) is generated instead, indicating that the allocation is questioned by the EPO (resp. USPTO). Resolution of the RHF takes place when the USPTO (resp. EPO) deletes the corresponding allocation [2]. The query language in Espacenet does not allow filtering on the level of invention versus additional information and/or the unreviewed/raise-hand status. For example, the query "cpc = H01L33/56" results in documents being classified in H01L 33/56 for invention information, for additional information and includes documents having the unreviewed/raise-hand status for this symbol. Internal analysis via EPODOC shows that for all the documents classified in the CPC with the symbol H01L 33/56, the symbol was allocated as invention information for only 34% of the documents and as additional information for 44%. For 18% of the documents the symbol H01L 33/56 has the unreviewed status and for 3% the raise-hand status.

5. Different classification practices between various patent offices might complicate the documentation coverage analysis. For this reason, it might be reasonable to do a sample check on patent families classified in the CPC to see the different IPC symbols provided by the national offices for the patent publications of each of the members. An example is shown in Fig. 7 for a patent family containing US2014174792, KR20140082292 and CN103897406. In the CPC all family members have the same symbols allocated, in this case subgroups of H01B3 and H05K1. The IPC symbols allocated by the USPTO are also subgroups of H01B3 and H05K1. However, the IPC symbols allocated by KIPO are subgroups of C08G and C08J (not visible in screenshot) and a subgroup of H05K1, i.e. H05K1/03 (which is different from the CPC symbol and the IPC symbol allocated by the USPTO). The IPC

Fig. 7. Screenshot from Espacenet showing the "INPADOC patent family" comprising US2014147792. The CPC symbols for the US, CN and KR publication are the same.

symbols allocated by SIPO are subgroups of B32B and a subgroup of H05K3. This makes the analysis on the documentation coverage of the CPC for the subgroups H01B3/30 or H05K1/185 more complicated because it is not clear if KIPO and SIPO systematically allocate the above-mentioned symbols to documents with a similar content as this patent family.

It is noted that the INPADOC patent family is defined more broadly than the simple patent family (mentioned in section 3.2). In the INPADOC patent family domestic application numbers are taken as additional connecting elements and documents having the same scope but lacking a common priority are included. However, in the case of the example shown in Fig. 7, the simple patent family is the same as the INPADOC patent family. In Espacenet, the simple patent family is shown under the link "also published as …" in the "Bibliographic data" page of a document (e.g. use the link "Bibliographic data" shown in Fig. 7, on the left hand side under US2014174792 (A1)). However, in the bibliographic data page only the IPC symbols of a single document are shown, not the IPC symbols of the other documents in the simple patent family. This makes a comparison of the IPC symbols somewhat more cumbersome.

4. Conclusions

In this study we have analysed the patent documentation coverage of the CPC in comparison with the coverage of the IPC. Patent families classified in the CPC have on average several family members whereas the patent families that are not classified in the CPC and are only classified in IPC typically contain only one patent document. For this reason, it is considered more relevant to compare the CPC patent documentation coverage to the IPC patent documentation coverage in terms of patent families.

Of all the patent families classified in the IPC only about one third is currently classified in the CPC. Most of the patent families that are not classified in the CPC are patent families including published patents applications filed at the three major Asian patent offices, i.e. the JPO, KIPO and SIPO. Since SIPO and KIPO recently joined the CPC the CPC patent documentation coverage is expected to significantly improve relative to the IPC coverage in the coming years.

We have demonstrated a method of comparing the CPC patent documentation coverage to the IPC patent documentation coverage down to IPC subgroup level. Such a comparison allows determining the importance of searching outside of the CPC for the concerned subgroup and how the search outside of the CPC can be done, e.g. if it is useful to search with the FI (File Index) and F-term classification. With this information, the efficiency of the search can be improved.

Acknowledgements

Disclaimer

Any opinions expressed in this article are those of the authors and not necessarily those of the European Patent Office.

References

[1] CPC Annual Report 2015 (http://www.cooperativepatentclassification.org/publications/AnnualReports/CpcAnnualReport2015.pdf).

[2] P. Held, Cooperative Patent Classification (CPC); Epoque Net http://www.cooperativepatentclassification.org/publications/CpcEpoqueNet.pdf (slide 4, explaining the status Confirmed, Unreviewed and Raise-Hand).

Bart Degroote, BE, patent examiner, Electrical and Electronic Technology, EPO The Hague. Obtained a PhD in physics from the University of Leuven. Prior to joining the EPO, he worked three years for Imec, a micro- and nano-electronics research center. He joined the EPO in 2006 and currently works in the field of printed circuits. He is a member of the Asia Patent Expert Group.