E

WIPO

IPC/SEM/98/2
ORIGINAL: English
DATE: November 20, 1998

WORLD INTELLECTUAL PROPERTY ORGANIZATION
GENEVA

ADVANCED SEMINAR ON THE INTERNATIONAL PATENT CLASSIFICATION (IPC)

Newport, United Kingdom, December 7 to 11, 1998

NATURAL LANGUAGE SYSTEM FOR ACCESSING THE IPC AND PROBLEMS IN THE PREPARATION OF AUTHENTIC VERSIONS OF THE CLASSIFICATION

Presentation by Mrs. Michèle Lyon, Chargée de mission,
Patent Department, National Institute of Industrial Property, Paris

LANGUAGE RELATED PROBLEMS IN THE INTERNATIONAL PATENT CLASSIFICATION AND SEARCH SYSTEM USING NATURAL LANGUAGE IN THE IPC

Introduction

1. The International Patent Classification, more commonly known as IPC, was officially created in 1971, when the Strasbourg Agreement was signed, and entered into force in october 7, 1975 in thirteen countries. It has been used for 23 years at least, by a continuously growing number of countries, but its real career has been much longer in fact, since some offices started to use it earlier. The French patent office, for instance, has allotted the IPC symbols since the late 1960s, as well as the German and the United Kingdom patent offices.

2. Since that time it has of course undergone a lot of modifications, either to take into account the technical progress and the amount of patent documents continuously added to the search files, or to improve the system and adapt it to the new needs of patent search.

3. Twenty-three years represent quite a long life for a search tool when considering the giant steps accomplished meanwhile by archive and retrieval techniques used in documentation along with the transformation of patent office policies prompted by electronic carriers and communication means. The new Standing Committee for Information Technologies (SCIT) has just been set up in WIPO and has decided to create a global information network between patent offices, this will undoubtedly have some consequences on the IPC.

4. In that context, it is time to assess the work done and to question the appropriateness of the IPC as a search tool for patents in the forthcoming years and possibly to adapt it in order to improve its efficiency and to facilitate use of this complex system for patent examiners and outside users.

5. This topic is very broad, hopefully the present Advanced Seminar will be able to discuss all the aspects involved, since it is really a multifacet question. Let's concentrate in the present talk on the linguistic issues in connection with the following aspects :

    - IPC and languages
    - The wording of the IPC : a crucial issue for the revision work
    - A revision procedure to meet the goals in English and French
    - Cib-Ln, a linguistic tool for an easier access to the IPC.

1. IPC and languages

1 - 1 Different versions of the IPC

6. To begin with, the IPC has been created as a bilingual tool : according to the article 3 (1) of the Strasbourg Agreement, which sets forth the IPC, "The classification shall be established in the English and French languages, both texts being equally authentic".

7. The English and French versions are strictly identical : not only is the layout of the IPC exactly the same in both authentic versions, with the same number of entries, but the words and phrases are carefully chosen (see below the revision procedure) to make sure the concept covered does not differ in any way.

8. Furthermore, the article 3 of the Strasbourg Agreement, in its second indent, refers as well to "Official texts" to be established in several languages by the International Bureau in consultation with interested governments. Presently there are five such texts. In addition to that, there are also "non official" (5?) versions of the IPC facilitating the access to the IPC in other languages.

9. The "official texts" of the IPC in other languages are elaborated by interested offices with the help of the International Bureau, so they are similar to the authentic versions and they can be relied upon to find relevant entries for a patent search.

10. This means that the user has the possibility to consult the version he likes, or rather the one he understands better, he will anyway find the appropriate symbol(s) to carry out his patent search on the subject matter he is interested in.

11. Issued in such a variety of languages, the IPC tends to get closer to the "universal value" claimed in the introductory part of the Strasbourg Agreement. However, the intended universality results as well from the possibility to overcome the barrier created by languages in the retrieval of patent documents.

12. Indeed, let's imagine our searcher referred to in paragraph 10 above if he had no IPC. He might use several domestic classifications, but this would just multiply the difficulties met when using (or revising) an international classification rather than solve them. He would then have to try to use words to describe the desired subject matter and to look for patent documents containing these words.

1 - 2 To use the IPC or to search with words ?

13. First of all, it should be noted that such a procedure would apply only to electronic carriers allowing word search (i.e. character coded documents), and people using paper would be able neither to make search files nor to search.

14. Incidentally, one has to remember that, according to § 52 of the Guide, an IPC symbol is allotted for an invention as a whole, and not for its constituent parts, which are nevertheless very often mentionned several times in the patent document. For instance, looking for documents describing an optical sensor through a request using only those two words will result in a considerable amount of documents which are not relevant, since a document containing those two words does not necessarily relate to that kind of invention, but might simply disclose a conveyor provided with any non described optical sensor. On the contrary, a search with the appropriate group symbol(s) of G01J for instance will allow to collect only the inventions relating to the desired optical sensor.

15. It is obvious to everyone that patent information has to be exhaustive in many cases. Bearing that in mind when searching, to select the appropriate words is really a difficult exercise, it is hardly possible not to forget any synonym, and to guess which phrases the patent applicant would have chosen is sometimes similar to gambling. It is unlikely that all people interested in patent information would be in a position to find the right answer, even in a single language, but difficulties get even worse when applying the process to several languages, which has to be done for most of patent searches.

16. It is undoubtedly much easier to read a well established IPC, even in a foreign language, than to imagine on one's own the appropriate words to use, which appears to be an endless query.

17. It does not mean of course that words will never be useful. A search using normal keywords may in some cases be helpful to find a "track" : having found a few relevant documents, it is convenient to use the classification symbols appearing on those documents to perform a thorough search. Or in a particular context, it can be very profitable to combine a search based on classification symbols with a search using words, because then, the domain is limited and the words allow to discriminate the documents. But in no way this can be considered as a universal solution.

18. Last but not least, as all members, imitated by a lot of non-members, of the Strasbourg Agreement allot IPC symbols to the patent documents they publish, it is clear that the user who carries out a search with IPC symbols in a given international documentation, will reach there any patent document which discloses the invention corresponding to the specified IPC symbols, no matter which language is used in the document.

19. Undoubtedly, such a possibility is essential and really gives the IPC its full "universal value" intended when the Strasbourg Agreement was signed, taking into account that any assessment of novelty or inventive step is always based on worldwide patent documentation. This advantage is unique, it gives the possibility to reach something relevant although unknown.

20. Those issues are certainly not new, they were assumedly considered before the IPC was established, and IPC specialists have been aware of them for a long time. The reason why attention has to be drawn once more to them is that they have to be assessed again in view of the new tools available at present, especially electronic documentation and search engines, bearing in mind the time consuming revision process and the complexity of the IPC for the user.

21. The relatively powerful search engines available nowadays put a strong temptation in searchers'way; but when they are used to search directly in the patent texts, search engines cannot solve the problems related to the range of languages used in patent documents or to the structure of a patent disclosure, which normally contains prior art description as well. A high number of occurrences of the target word(s) cannot be systematically interpreted as a positive indication of the document relevance. The combinations of words which can be included in a request sometimes increase the number of uncertainties and the target remains difficult to define. In short, those new technologies are disappointing when applied to patent texts and they narrow the linguistic possibilities instead of widening them.

22. So, it appears that giving up the IPC now would mean loosing the famous "universal value" and going backwards. From the linguistic point of view, there is still no satisfactory way out apart from the IPC. This issue might have to be considered again from time to time in the future, since technology is improving quickly, but in any case, the search tool(s) to be chosen will always have to be of "universal value".

23. If the IPC remains the only acceptable tool, it should certainly not remain static. On the contrary the IPC should be developped and adapted both to the new technical subjects disclosed and to the kind of searches performed to day.

24. The problems related to the structure of the IPC will be touched upon later on during this seminar, but from the linguistic point of view in particular, the level of quality should at least be maintained and even improved.

2. The wording of the IPC : a crucial issue for the revision work

25. The study of IPC revision projects is in general carried out very carefully since any remaining mistake or inaccuracy may have enormous consequences in case of document misplacement or of misoriented search. A great deal of efforts are invested in the elaboration of the IPC to ensure a classifying of patent documents as efficient and consistent as possible all over the world.

26. This is perfectly in conformity with the Strasbourg Agreement which in its article 5-3-ii assigns to the IPC Committee of Experts the task to "address recommendations to the countries of the [IPC Union] for the purpose of facilitating the use of the Classification and promoting its uniform application". This policy statement should of course apply to the Committee of Experts itself when elaborating the IPC.

27. Accurate language is certainly of a great help to enhance the uniform application the IPC Committee of Experts has to promote. The choice of the proper words or phrases for the classification entries is essential and it is really important to discuss them very carefully. They should be understood the same way by all readers to ensure the correct and consistent placement of the documents in IPC entries by the examiners, as well as the retrieval of those documents. The time spent to discuss the wording looks like a kind of investment to simplify and enhance the search.

28. The external users, and sometimes even the internal ones, are quite often complaining that the IPC does not use the common language, that the simple words are not enough used in IPC. It is true that, when created or revised, IPC entries have to provide for yet unknown subject matter to be catered for ; therefore it is often necessary to insert periphrasis covering all unpredictable inventions which would be otherwise excluded by words with a more restrictive meaning. Possible solutions to this problem are to increase the number of precise examples in the wording or to enrich the catchword index, a real help for electronic searching, as explained below (see paragraph 64).

29. At the same time, any extensive interpretation of the IPC wording should be carefully avoided, the subject matter has to be clearly limited to what is intended.

30. Another reason to pay full attention to the wording of IPC entries is that automatised processing of the established text is expected to take place in the future, even if we don’t know yet to which extent. The only way to ensure reliability of automatised systems is to be consistent and rigourous.

31. In that context, the wording of the IPC entries should be even more carefully drafted, and perhaps the efforts should be increased in that respect when revising the IPC. The lists of words or phrases already used in IPC is easy to consult with electronic carriers such as the CD ROM Ipc:Class, and may be the stress could be put on harmonization in a more intensive way.

32. Needless to say, all the above considerations have to apply to both authentic versions. However, once the first authentic version has been modified in an IPC revision project, the efforts involved to revise the second one are far less than those put into the changes of the first version, provided that the revision procedure allows to take advantage of the input during the first phase.

3. A revision procedure to meet the goals in English and French

33. It is probably useful to describe this process here, although it is part of the present IPC revision procedure, because the offices working exclusively in English might not be thoroughly aware of it, and because the adquired experience has to be taken into account when exploring new ways.

34. Once some IPC modifications have been adopted in one of the two authentic versions in the course of a project, a translation process is initiated to prepare the corresponding version in the other language. From a practical point of view, looking back to the revision work as carried out during the last revision periods, it appears that the translation process has been performed from English to French in most cases.

35. Every step in the translation process could be applicable whatever the language is, English or French, of the IPC modifications adopted first. However, to make things simpler, let's take the most frequent exemple where the IPC modifications are first adopted in English and then translated into French.

36. The procedure is well established, although not accurately described in the instructions for IPC revision appearing in the WIPO Handbook on industrial property information and documentation.

37. French speaking offices, that is to say European, Swiss and French patent offices, usually share the work and distribute the projects amongst themselves to establish the french version of modifications already adopted in English, sometimes joined by the Canadian patent office.

38. One of the french speaking offices prepares a provisional french version and submits the result to all other offices participating in the revision work and to the International Bureau; the other offices perform a critical examination (in a positive acceptation of the phrase!) of the proposal and send their comments to the author on

    - clerical mistakes (numbers, dots, spelling or typing mistakes…);
    - correctness, accuracy of the resulting text, possible ambiguïties;
    - consistency in words and phrases used within the revision project and in other parts of the IPC.

39. The provisional French version is very often sent back with a large number of suggested improvements, not because the originating office has done a bad job, but because it is a difficult exercise. The quality level requested has to be absolutely identical to the one of the English version, to establish both texts as equally authentic.

40. The office responsible for the French version in a project thoroughly studies every suggestion and introduces the ones deemed appropriate into the text. The result is presented to the Working Group in charge of the adoption.

41. The preparation of the French version constitutes in fact a systematical verification of the English part already adopted by the technical experts : at that point, it is not unusual to find some remaining mistakes in the English version. Moreover, the translator sometimes comes across ambiguïties in the initial text which have not been seen during the discussion on the technical aspects. He has then to ask for clarification in order to establish the text in French and a feed-back takes place to clarify as well the English version.

42. One has to remember that authentic versions of the IPC are used all over the world by classifiers or searchers whose mothertongue is neither English nor French. The translation into French can be considered as a test for aiming at consistency in IPC use (see paragraph 27 above).

43. Furthermore, the authentic versions serve as a starting point to establish many other IPC versions in various languages at a time when it is generally much more difficult, if not impossible, to correct the text before the following edition.

44. Not only is the translation process intended to create both authentic versions, but it certainly enhances as well the quality of the version first established. Let's notice that, since systematic checking is anyway carried out by the French speaking offices in the frame of the translation process, this additional degree of quality for the English version is in fact reached with a modest contribution from the English speaking offices which are called in only if a problem arises.

45. Moreover, the above mentionned analysis of the English version for translation takes of course place at the office and does not extend at all the discussion during the meetings.

46. Speaking about meetings, it should be noted that the elaboration of the French version has little influence on the total time spent on revision projects : most of translation problems are generally solved in advance by correspondance between French speaking offices and there are few exceptions ; the adoption of the French version is then restricted to endorse the text proposed by the office responsible for it without any lengthy discussion and is just a formality. It is well known that in a meeting, projects where the only pending task is the adoption of the French text can be completed in a very short time at the beginning of the meeting, allowing the chairman to quickly improve his (her) score in relation with the number of « discussed » projects !

47. For all those reasons, it is essential to keep the parallel elaboration of both authentic versions which strengthen one another, since the benefit for non french speaking offices is considerable in regard to the little complication involved and short time spent. The adoption of the second version has anyway to take place sooner or later.

48. It is clear that a delayed adoption of the French version would be harmful to the quality of both versions, since a feed-back at the level of the Committee of Experts or of a group without the technical experts would not have the same impact.

49. English and French versions being equally authentic implies an exact similarity between both. Such a result can only be achieved if the translation process occurs within the technical study of revision projects, when all the issues involved are still present in the mind of the technical experts, among which are the translators. As an office which has always actively participated to the elaboration of the French version and has gained a solid expertise , INPI might not be in a position to do the job if delayed, because it would be more difficult, take longer and lead to a result far less reliable.

50. Similarly, it would make the work harder if the technical meetings were to be held only in a single language. It is true that, in the past, "sub-bodies" dealing with a very precise subject matter, in general one revision project, have been very successful because the participants were the experts in the discussed technical field; the language spoken in the last four or five sub-bodies was English, and it was the only way to cope with the task. As an exceptional procedure, it has been very useful and could certainly be repeated. Nevertheless, it is very difficult to imagine a french version of the IPC

51. discussed only (or even mainly) in English ! To build a real French IPC version , the french language has to be used during the technical discussion of revision projects, otherwise part of the problems peculiar to the french language will remain hidden and unsolved.

52. It should be noted that the article 3 (1) of the Strasbourg Agreement sets forth an obligation to reach a particular result, i. e. to build two versions equally authentic. Consequently, the Committee of experts has to implement the appropriate procedures in order to obtain this compulsory result. The French office would strongly object to a deviation from this policy.

53. It is worth stating that the French version is not the French offices problem only : again, if French speaking offices, with a « second choice IPC », were not in a position to allot the right IPC symbols, IPC would no longer meet its goal as universal system designed to retrieve all patent documents.

54. The possible changes to be introduced in the IPC should at least maintain the level reached so far and the same applies to the revision procedure, which has to contribute towards quality and consistency. It may well be that, in the near future, no revolution can be expected in the linguistic field to boost or replace the IPC, and that the combination of various available tools is a way out. In that respect, it appears that another field to explore as well is to facilitate the use of the IPC and the access to it for all kinds of users.

55. For that purpose, the French office has developped a linguistic tool, CIB-LN (standing for "**CIB** en **L**angage **N**aturel" or Natural Language IPC"), the purpose of which is to help any user interested in a particular subject matter to find the corresponding IPC symbols.

4. CIB-LN, a linguistic tool for an easier access to the IPC

4 - 1 Goals and historical background

55. As it will clearly appear later on, this tool is not intended at all to replace IPC experts, but rather to assist users lacking experience who find themselves at a loss when looking for entries relevant to a particular subject in the IPC, and to increase public awareness about patent information.

56. The need for such a tool has probably existed since the birth of the IPC, but it increased dramatically at the French office when patent search was made available on line for a large public. The IPC appeared then to be a problem to occasional users not very familiar with it. Nevertheless, no alternative search means was available at hand (as discussed above), and the patent documents (at least the french and european ones) had been classified only according to the IPC for years, so that it was impossible to get rid of the IPC.

57. After a while, a new service was provided by INPI on videotext : every user could put a question on its "Minitel" asking where a particular subject matter was covered in IPC. The requests were collected every day and a few examiners in charge of this service indicated the relevant IPC symbols in the user's"letter box". But it was a rather heavy system, since the first request was not always clear, and a further dialog between the user and the examiner somewhat cumbersome.

58. The idea came through of an automatic system where the user could type its request on a Minitel or a computer and get answers at once. The user would not be expected to use the IPC language, but reather to word his query in natural language, and the system would return the relevant IPC symbols, which could possibly be directly used for an on line data base interrogation.

59. At the beginning, it was mostly a dream, and no one knew whether it was feasible or not. A company specialised in linguistic engineering, Erli, was approached in 1993, and after they had carefully studied the IPC and its rules, the decision was taken to build a prototype.

60. The prototype was limited first to subclass titles and then extended to some main groups in sections A and B. Although the results were incomplete due to the limitations of the prototype, it was felt that they were encouraging enough, and that it was worth going on and establishing a system enabling queries in natural language for all parts of the IPC. The company Erli was entrusted with the job of building the system, and the overwhelming development of the Internet required of course to design it for use on the Web.

4 - 2 Realization of the system

61. A lot of efforts and enthusiasm have been invested in that project within INPI. A team comprised of IPC specialists, data base specialists, computer experts was set up in the office in order to give Erli all the necessary information, to define the goals and give some advice on technical choices to be made and finally to check whether the results returned were satisfactory. In particular, the IPC specialists had to ensure that IPC rules were taken into account and that queries got accurate and complete answers.

62. Erli processed in depth the text of the IPC intended to stand behind the scene of the system, in order especially to include the operating rules of the IPC, and in particular the hierarchical structure. The text of the hierarchically superior entries has been included in every entry in order to make the system work as a human reader taking

into account the implicit wording of every group and the number of dots. Along with this move downwards, the words of the dependant subgroups have been inserted up in every main group to give a representation of its whole contents, since the system is designed to return answers not below the main group level. Those treatments are of course invisible for the user who can eventually read the true authentic text of the IPC similar to the one appearing in the books!

63. Apart from that, Erli has used its own dictionaries, specially enriched for this application, and implemented its search engines to perform linguistic analysis of the text of the IPC as well as of the query. The exact process applied is based on rather complex linguistic engineering techniques, but it takes into account a lot of parameters such as the presence of the search phrase in the IPC entries, total or partial synonymy, relative weight of words, and so on…The query is also analysed and commonplace words such as "apparatus", "method" are weakened enough not to attract all the IPC entries containing those words.

64. The well spread use of periphrasis in the IPC, as mentionned earlier, is really a challenge for that kind of system because they do not necessarily contain the words to be looked for and because they make sentences longer and much more difficult to analyze. The solution adopted is an "over indexation" (with a meaning different from the one used in the IPC) and is mainly based on

- the catchword index

  when the catchwords pointing to an IPC entry do not appear in the IPC, it has been artificially added to the wording of the corresponding entry. It is the reason why introducing new catchwords would certainly be helpful for the user.

- indexing schemes

  in some cases, the indexing schemes contain concepts and vocabulary very useful for queries. Some indexing schemes have been "introduced behind" IPC entries with the necessary processing to reflect the essence of such codes.

- when appropriate, subclass indexes have been used as well.

4 - 3 Testing of the system

65. A batch of about 350 queries from outside users having really ordered searches to INPI has been used, and the IPC team allotted all the possible relevant IPC symbols (with two levels of relevance if appropriate) for every query.

66. This corpus of 350 queries along with the correct IPC symbols intellectually allotted have been used to test the system. Every time a series of changes were introduced, a calculation was made to assess various ratios, for instance :

- Recall ratio = the number of correct answers obtained / number of expected answers;

- Pertinence ratio = the number of correct answers obtained / number of obtained answers;

At the end of the project, the recall ratio was about 79% and 55% of the expected answers appeared among the first 20 answers given by the system.

4 - 4 Boundaries and assessment of the system

67. As stated above, the fact that most of the users'questions are rather general and that it is very often not possible to give a precise (and safe!) answer at the subgroup level has been taken into account : the system gives answers no lower than the main group level, and the user has to go to the place in the text of the IPC to see the subgroups in their hierarchical position; he can also navigate through the IPC and display various combinations of the hierarchical layout, or even directly ask for an IPC symbol if he knows it.

68. The system can be operated for most parts of the IPC. However, in the chemical field, the language is so specific that it cannot be handled by linguistic engineering, not to speak about the last place rule. It has been felt wiser not to developp the access to chemical groups through chemical names or descriptions, and to limit the retrieval of chemical groups to those which can be reached with "ordinary" language, such as "medicines", "coatings", "micro-organisms" etc…In any case, all the subgroups, even in chemistry, can be displayed in the navigation mode.

69. The system has been launched in last september on the INPI web site, as part of the French Industrial Property Digital Library, and it is free of charge. The answers are given in the order of pertinence. The user, having selected the appropriate symbols in the very text of the IPC, can access the corresponding FR, EP and PCT patent applications valid in France

70. At the time this text is written, no real feedback has been received yet and it is to early to assess its success. It is worth saying however that it has been favourably welcome when presented to some people working with IPC.

71. By no means CIB-LN is intended to be a fully automatic system, the searcher has eventually to choose the appropriate entry in the IPC itself. It is a linguistic system, it does not have any "technical expertise". This is partly balanced by the dictionaries and the use of synonyms when available, but it is designed to help users not familiar with the IPC, not to replace experienced examiners.

72. INPI is aware of some weaknesses of the system, but is convinced that the benefits are worth accepting them. Furthermore, the system is expected to continuously improve and undergo necessary adaptations through the maintenance process. The results obtained in the course of operation will be carefully studied, especially to detect the questions with poor or no answers. The solutions will mainly consist of adding new words not present in the IPC, either as synonyms to existing words or as additional catchwords pointing to specific IPC entries. Along with that, it is possible to adjust the weight of the words and of the relationship between partial synonyms. It will of course take some time.

73. So far, the Cib-Ln system exists only in French. However, the structure of the systems allows to build the equivalent in other languages without starting again from the very beginning. A presentation of the system will be given during the Advanced IPC Seminar.

Conclusion

74. The seventh edition of the IPC is about to be issued, which means that, over the years, the IPC has somehow proved to be an efficient and easily adaptable patent search tool for specialists, and yet no alternative satisfactory global search means is available. To day, electronic means associated with linguistic engineering offer an opportunity to widen its field of use and to allow a larger number of users to access it as well, especially in the framework of the Global Information Network. It would be a pity not to grasp such a possibility, provided that the level of quality is maintained. In that context, the criteria of "universal value" would be even better fulfilled. Along with the unavoidable technical adaptations to be implemented in the structure of the IPC, the necessary efforts should be input in the revision work to keep it alive in both authentic versions.

[End of document]