system. Many decisions are yet to be made regarding these code modifications, but it is possible to list some that are definitely in the offing, and others which could greatly improve retrieval from WPI. These are summarized in Figure 9.

A select list of about 2000 common chemicals will be assigned registry numbers. They will be registered as starting materials, as well as when they are products or are otherwise used, and their function will be described by a role indicator. While it would be best to have all starting materials coded, the registration of common starting materials should take care of the most urgent needs. For a while Derwent talked of creating its own registry numbers, but now it appears that they will use CAS registry numbers, a wise decision in light of the growing use of these numbers in the literature.

At one point Derwent contemplated specifically coding each compound which was claimed or which appeared in an example, in addition to any Markush structures claimed. This would have been extremely valuable in improving precision, but apparently has been shelved. New overcoding rules which reduce the amount of overcoding on a given input record should help somewhat, but ideally each individual compound coded should be coded separately from all other compounds. The American Petroleum Institute's data bases have shown how each separate chemical entity can be coded, with its terms linked, so that they can be kept apart from other compounds in the document.[6]

Freedom from IBM card input allowed the introduction of Ring Index numbers, and an extension of this principle was planned to allow the specific coding of all elements, ring substitution patterns, and added functional groups. Hopefully Derwent will extend this principle in a general restructuring of the multipunch code format. Various proposals have been made in this area, but this point in particular is still very much up in the air.

Molecular formula can be a vital data element in solidifying a structural retrieval system. There is perhaps no simpler way of enhancing the retrieval from CPI (especially for inorganic compounds, but broadly for all compounds) than the coding of molecular formulas. A few multipunches plus a molecular formula would constitute a powerful retrieval tool. At present Derwent has made no commitment regarding the possible inclusion of molecular formulas in its indexing.

Other steps that would improve various aspects of subject retrieval would be the coding of chemicals that presently are not placed in Sections B, C, or E; more liberal use of added keywords; entering of all international patent classes assigned to a patent, even if they are related to previously assigned IPCs; and amplification of Derwent's catalyst codes, which presently lump together elements from different groups of the periodic table, a most unfortunate circumstance.

WPI is the world's most all-encompassing file of patent information. Even with its present retrieval limitations it is an invaluable information tool. Elimination of those limitations could provide us with an extraordinary patent information resource for the future.

REFERENCES AND NOTES

(1) The story of Wöhler's letter, written in the name of S.C.H. Windler, is recounted in CHEMTECH **1978**, *8*, No. 12, 757.
(2) Duffey, M. M. "Searching Foreign Patents", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 126-30.
(3) Kaback, S. M. "A User's Experience with the Derwent Patent Files", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 143-8.
(4) Kaback, S. M. "Retrieving Patent Information Online", *ONLINE* **1978**, *2*, No. 1, 16-25.
(5) Several additional studies of retrieval from Derwent files may be found in the Proceedings, International Patents Conference, Stratford-upon-Avon, England, April 12-14, 1978; Derwent Publications Ltd., London.
(6) Kaback, S. M.; Landsberg, K.; Girard, A. "APILIT and APIPAT: Petroleum Information Online", *Database* **1978**, *1*, No. 2, 46-67.

The Approach of the United States Patent and Trademark Office to Finding Prior Art†

ALFRED C. MARMOR

Patent and Trademark Office, Washington, D.C. 20231

Received May 14, 1979

The foundation for effective retrieval of technical information by the U.S. Patent and Trademark Office (PTO) continues to be the U.S. Patent Classification System. Utilizing this system, the PTO currently maintains over 22 million documents, including U.S. patents, foreign patents and applications, and a variety of nonpatent technical disclosures in a viable file array of approximately 100 000 distinct, defined subdivisions. Providing prior art searches requires the maintenance and updating of a very large search file and a complex classification system. Continually developing computer-based systems are permitting the PTO to take an increasingly sophisticated approach to search file management. These systems, as well as newly developing information systems technology, will provide the capability to find prior art more readily and more precisely from an ever-growing search file.

The United States Patent and Trademark Office (PTO) must discharge three primary functions. The first is to examine applications for patents and, after examination, to grant or refuse to grant a patent. The second is to examine applications for trademarks and to register or refuse to register the trademark. The third is to collect, classify, and disseminate technological information disclosed in patents and related technical documents. This paper is concerned with the first and third functions.

The requirements for meeting the obligations under the first function, i.e., patent examination, make it necessary for the examiner to determine whether the invention meets the stat-

† Presented in the symposium on "International Aspects of Technical Information Retrieval", Division of Chemical Information, 177th National Meeting of the American Chemical Society, Honolulu, Hawaii, April 2, 1979.

utory criteria of novelty and unobviousness as set forth in Title 35 of the United States Code. A component of this determination is consideration of prior art.

Prior art is, essentially, the total body of publicly available recorded information and knowledge in every field of science, engineering, and technology which has preceded a claimed invention. Clearly, any PTO decision to grant or refuse to grant a patent, based on the statutory requirements, would be impossible unless the necessary tools are made available to the examiner. The primary tool, of course, is the classified patent search file.

THE PATENT SEARCH FILE

The PTO has assembled and maintains one of the largest collections of technological literature in the world. This unique file of prior art contains 22 million documents classified and distributed among nearly 100 000 discrete subdivisions of technology and is growing at an annual rate of about 250 000 new U.S. patents and some 300 000 new foreign patents. Because of its nature and the attention that has been paid to it over the years, this file represents a unique national resource. As such, the PTO recognizes its heavy responsibilities not only to preserve, maintain, and improve this resource, but to maximize its use by the public for the greatest possible national benefit. Accordingly, techniques to locate and retrieve, from this mass of literature, information relevant to a particular field of technology are under continuous development.

Of course, the ideal situation would make it practical for the examiner to retrieve every pertinent document from any source and apply it against the patent application being considered. Realistically, however, this objective must be achieved on a cost-effective basis within the constraints of time and budget. In addition to the examiner's use of the file, there also exists the public's use in determining the state of the art and identifying potential products for the American and international consumer.

The classified prior art file (patent search file) is composed of three types of documents: first, all U.S. patents; second, foreign patents; and third, nonpatent technological literature. Although more than four million U.S. patents have been issued since the first patent was granted on July 31, 1790, there are now in the file approximately 12.5 million U.S. patents. This nearly threefold expansion of the U.S. patent portion of the file is because each patent may contain information which fits more than one specific category, thereby requiring a copy be placed in each of the relevant categories.

The PTO receives copies of foreign patents as a result of patent exchange agreements with foreign patent offices. The classified file contains 9.5 million foreign patents about evenly distributed between English and non-English language patents.

Each year approximately 600 000 foreign patents are published worldwide, many of which are duplicative of each other. The placing of all duplicate foreign patents into the file would excessively expand the file without sufficient concomitant benefit. Moreover, many of the foreign patents are in languages not readily understood by most examiners, for instance, patents printed exclusively in Far Eastern, Indian, or Arabic languages. Patents printed in Romance or Germanic languages ordinarily present lesser difficulty but would not be as useful generally as those in the English language.

To improve general readability of the file, English language abstracts are being associated with foreign language patents. Ordinarily, the abstract is added to the first published patent of a family of patents. For example, if the first of a family is a Japanese patent, the English abstract is affixed to a copy of the patent and the copy then placed into the classified file. If, later on, an English language counterpart of the Japanese patent is received, it may be used to replace the Japanese patent. If the later-received English language foreign patent has an effective date subsequent to the Japanese patent, the Japanese patent will be retained in the file as providing an earlier date for optimum prior art purposes. Accordingly, the classified file should contain no more than two copies of each foreign patent while providing the best prior art available worldwide.

The third component of the classified prior art file, the nonpatent literature, is handled in two ways. First, articles from technical journals, which are deemed of importance for the search file, are classified by a contractor, according to the U.S. Patent Classification System. Secondly, several hundred technical journals are circulated to the patent examiners, through the facilities of the PTO Scientific Library, who may and often do select all or parts of technical articles for inclusion in the classified search file.

The effort to include more extensive nonpatent literature documentation into the search file stems not only from a need to provide maximum prior art to patent examiners but also to meet the minimum documentation requirements under the Patent Cooperation Treaty. This treaty, of which the United States is a member, came into effect in 1978. Under the treaty a person residing in any signatory country may file an application in that country and later obtain patents in other signatory countries based upon the same application and without filing separate applications in each of the other countries.

THE U.S. CLASSIFICATION SYSTEM

The patents in the patent search file have been categorized according to the U.S. Patent Classification System. This system, one of the most sophisticated known for classifying technological information, provides for a systematic arrangement, in separate, well-defined categories, of the information contained in U.S. patents. Exhaustive of all subject matter which is patentable, the system allows reasonably short and complete searches to be made by examiners in their determinations of patentability, and by inventors and the public-at-large in their preexamination searches, as well as in state of the art, validity, or infringement determinations. The U.S. Patent Classification system, initiated by the PTO around 1900, has been continually revised and updated since that time. New categories are constantly being added to accommodate new technologies and to further refine the classification of existing technologies. Currently, there exist some 350 broad technological categories called "classes" and some 100 000 specific categories called "subclasses".

Early U.S. patent classification was largely industry oriented. However, as the patent system developed, it became increasingly evident that this approach was not satisfactorily meeting contemporary needs; e.g., the classification did not follow the same path as legal doctrines such as the notion of "functional equivalence". As a result, the system was revamped and the concept of "proximate function" was adopted as a primary basis for patent classification. Under this concept, devices or processes that inherently achieve similar results are classified together. Thus, all cutting machines are arranged together, be they for cutting wood, metal, or meat since they share the "proximate function" of cutting.

In accordance with this concept, and through analysis of the subject matter of each U.S. patent, broad categories (classes) and more detailed subdivisions (subclasses) of the subject matter have been created and patents collected into each appropriate category. Each of these classes and subclasses is titled. For each class there is a schedule which lists, in an ordered outline form, the title of each subclass therein, arranged to show the hierarchical relationships among subclasses (see the below reference to the manual). This type of

arrangement aids both in determining where documents should be classified and in limiting the area of search for well-defined features.

REFERENCE TOOLS

Highly effective and relatively easy-to-use publications are provided and maintained by the PTO to aid in the use of the U.S. Patent Classification system. These publications are: (1) *Index to U.S. Patent Classification* (the "Index") (2) the *Manual of Classification* (the "Manual") and (3) the *Classification Definitions* (the "Definitions").

(1) The Index comprises an extensive alphabetical listing of subject matter headings or descriptions intended to serve as an initial means of entry into the classification system. Each description is tied to specific citations of class(es) and subclass(es) where documents relevant to that description may be found.

(2) The Manual which includes the schedule for each class of invention is the key publication required for a search in the patent file. By using the Manual one may determine which class and subclass to search to find the desired subject matter.

(3) The Definitions augment the necessarily brief titles for each class and subclass appearing in the Manual by explaining each of the titles in a detailed statement setting forth the limits of the subject matter to be included thereunder. Key features of the Definitions are search notes which help direct searchers to all areas in the search file where relevant information may be located. Through such notes the definitions serve to obviate the need for overly extensive cross referencing.

By combined, systematic use of the Index, the Manual, and Definitions, one is able to locate any patent or other document that has been placed in the patent search file.

Despite the efforts of the PTO at keeping reference tools up to date, the updating cannot always reflect all aspects of a rapidly advancing technology. Being deeply aware of this problem, the PTO is constantly seeking means to improve access to the documents in the file. One such effort involves an experimetal on-line computer-based searching system for textual searching of all words in the Index, the Manual, and a set of augmented patent titles of U.S. patents which were issued in 1975 and 1976. Searching may be performed by free choice of terms and phrases using a portable desk-top terminal with telephone connection to the computer.

An immediate objective of the system is to provide subject matter access to the patent search file for those who are not trained in the use of the U.S. Patent Classification System or who are not familar with the given technological area. At the same time, it provides increased capability for subject matter searching of the patent file by patent examiners and patent attorneys.

Searches made on the Patent Title File identify subclasses which contain more recent technological advances or technology which may not have been made explicit in the Manual or Index. Topics such as "seismic prospecting" and "ink jet printing" may be found with high frequency in the patent titles and yet neither phrase can be found in the Manual. Common technological vocabulary is also more likely to be found in the Patent Title File than in the Manual where a concept is usually described more generically.

FILE MANAGEMENT

It could reasonably be questioned as to how soon the sheer volume of the search file will become so great as to make it no longer feasible to continue maintaining and using the file. It is unlikely, for example, that all of the documents maintained in the file are being used frequently enough to pay their way. But which ones? With the passage of time, a static search file could become technologically obsolete. How should PTO resources be utilized to minimize such obsolescence? Present methods of utilizing the paper search file inherently result in a continued degradation of file integrity. What can be done to maintain an acceptable level of integrity? It may not be cost-effective to maintain the entire file in paper copy. But which parts of the file should be converted, and to what form? Currently, the PTO is developing techniques for answering such questions.

The development of techniques for improved file usage is the heart of the Patent Documentation Organization (Documentation). Through a computerized data base compiled by Documentation, the PTO is able to maintain records of all patent documents and the subclasses in which they appear in the U.S. Patent Classification System, thus providing an essential inventory of file content.

The Documentation data base will also enable the study of the utilization of documents contained in the file. Analysis of search activity patterns, in conjunction with analysis of document citation frequencies, offers tremendous potential for learning more about the value of documents of differing ages, types, and countries. This knowledge permits a high degree of sophistication and precision in search file management and in the determination of future file content.

Ideally, this study and analysis will permit the PTO to perhaps purge some documents, such as older ones, from the search file or, as a minimum, store them in a different medium. However, the patent statutes require that the invention for which a patent is sought must be new and unobvious over all available prior art. Therefore, in order to purge documents from the file, legislation permitting such purging would be required.

In terms of total resources, personnel, and budget, the current major thrust for improving the search efficiency and quality of the file is reclassification. The U.S. Patent Classification System, like all other dynamic classification systems, must continually be updated to provide for new technologies and changes in existing ones. Reclassification is a complex process both logistically and intellectually and involves over 300 professional, technical, and clerical personnel on a full-time basis. Approximately 4% of the file is reclassified annually.

Projects for reclassification have to be selected carefully if resources are to be concentrated in those areas of the file most in need of improvement. Therefore, another important use of the Documentation data base is in reclassification project selection. By using the data base, the PTO is able to follow and monitor file growth. It has been found, for example, that on a short-term basis of a year or two, an average of 30% of the patents granted relate to only 10% of the subclasses making up the patent search file. This concentration of file activity is continuously changing in response to technological advancements and is a reflection of needs related to current public interests, such as ecology and energy.

To identify the areas most frequently searched and inadequately classified, a project selection technique has been developed which combines the professional judgment of classifiers and examiners and the objective data obtained from the computer record as to search activity and file growth. When any portion of the search file is to be reclassified, it is first subjected to an integrity check. After new classifications are determined for each patent, documents are converted to the new classification arrangement by a combination of clerical and computer processing.[2]

The Documentation data base is additionally utilized in the maintenance of an acceptable level of file integrity, thus improving the potential for a quality search. A lack of file integrity exists in the paper search file when a document is not present in a subclass at the time the subclass is searched. To some extent, some temporary instances of loss of file in-

tegrity are inherent in the use of the file, since documents are continually taken out and replaced in the examining process. This temporary removal is considered to be unavoidable, but the absence of documents for any other reason is considered unacceptable.

To minimize this problem, a file integrity operation has been established within the Patent Documentation Organization. This operation is designed to ascertain those documents unacceptably missing from the file and to replace them. In addition the operation provides feedback to permit upgrading of the data base itself. The result is both a better search file and a more accurate computerized base of data about the file.

Although the preceeding discussion focuses primarily on the content and use of the patent search file as a manual access paper file, it is not intended to be interpreted as an indication that the PTO has accepted as inevitable the continuance of the file in this form. Alternatives to paper as well as alternatives to conventional classification systems as the basis for search are currently being investigated.

COMPUTER CONTROLLED MICROFORM SEARCH SYSTEM

To help answer questions associated with the use of coordinate indexing systems, file integrity, and the ability of examiners to search documents contained on microfilm, the PTO has undertaken two experiments in the field of computer-controlled microform. The first combines the searching of a coordinate index computer record and the displaying of documents identified in the search. The second is directed toward the simulation of what an examiner does while searching a manual-access classified file. These experiments involve the use of a Computer Controlled Microform Search System (CCMSS).

The CCMSS is an on-line computer terminal system capable of (1) searching a computer record of index terms relating to the patents and other documents stored in the system, and (2) displaying, within a few seconds, images of the documents relevant to a specific query. The system consists of a number of terminals all of which are under the control of a single minicomputer.

Each terminal consists of a microfilm display device, a data scan, a keyboard, and a carousel/selector for storing and selecting microfiche. Some terminals also are equipped to make paper prints. Each terminal can store and display some 200 000 images or approximately 30 000 U.S. patents. The keyboard and data scan are used to communicate with the minicomputer system in which the records of indexing information for the patents are stored and processed.[3]

A number of factors are being studied, analyzed, and evaluated as a result of these two experiments. These factors include: (1) the effect that file integrity, inherent in such a system as CCMSS, will have on the quality of patent search; (2) the ability of examiners to search using microfilm as opposed to paper; (3) the effectiveness of coordinate indexing systems as alternatives to conventional classification systems; and (4) the impact of such systems on examiner search time.

FULL TEXT SEARCH

Looking further into the future, the PTO is not unaware of the possibilities offered by full text searching. The keying of patents in machine-readable form was initiated in the PTO in 1970 with the express objective of having a data base for computerized searching of the full text of patents. It was recognized in 1970 that while full text searching was beyond the state of the art at that time, it would likely come into being in the future; thus, in order to preclude a huge data conversion effort, the keying process was initiated. Generally speaking, much of the cost for implementing a full text search system is associated with the conversion of the text to machine-readable form. The PTO is, therefore, in a favorable position in that over 500 000 patents have already been converted and all issuing patents in the future will be converted to machine-readable form.

Full text search systems which have been implemented to date are generally used to search a wide-ranging corpus of disparate art, more often than not in a nontechnical area. There has been very little experience by anybody in full text search of homogeneous technical art consisting of thousands of documents, all dealing with the same narrow technical discipline, as found in patent searching. Significant resources will be required, therefore, to develop a full text system for patent searching.

Given the current state of government budgets, it is unlikely that these resources will be forthcoming anytime soon. However, in anticipation of the role of full text systems for searching prior art in the future, the PTO is, as a minimum, maintaining an awareness of the state of the art.

REFERENCES AND NOTES

(1) The full scope of the U.S. Patent Classification System goes beyond the scope of this paper; however, details may be found elsewhere in the literature. See U.S. Department of Commerce, "Development and Use of Patent Classification Systems", U.S. Government Printing Office, Washington, D.C., 1966.
(2) For further discussion of the classification process in reclassification projects see Kendall J. Dood, "The U.S. Patent Classification System", *IEEE Trans. Prof. Commun*, PC-22, 95-100 (1979).
(3) For additional information on the application of CCMSS to patent searching see P. M. McDonnell, "Computer-Controlled Microform Search System for Patent Searching", *J. Pat. Off. Soc.*, **59**, 175–179 (1977).