Anne M. Carpenter, Meriel Jones, Charles Oppenheim
Centre for Infonnation Science, The City University, London

Retrieval Tests on Five Classification Schemes

Studies on Patent Classification Systems II(1)

Carpenter, A. M., Jones, M., Oppenheim, Ch.: **Retrieval tests on five classification** schemes. Studies on Patent Classification Systems II(1).
In: Intern. Classificat. 5 (1978) No.2, p. 73–80
The five schemes tested in the fields of inorganic chemistry and biochemistry were the following: the UK Patent Classification, the US Patent Classification, the International Patent Classification, the Derwent Manual Code and the Derwent Punch Code. Description of the methodology followed and presentation of the results from the 15 and 22 questions set. Among others the results show that recall was correlated positively with precision and that the Derwent codes gave substantially higher recall than the classification schemes. Details of the findings are discussed including also the role of misprints in patent office publications. I.C.

1. Background

Many papers have appeared in the literature in which the retrieval performance of information retrieval systems is assessed. The most widely used measures for such an assessment are recall and precision, which have the advantage of being easy to calculate and which appear to have encountered no more, or less criticism than other measures of retrieval performance (2,3). The general procedure for a test is to collect a number of documents on one subject, and these are then indexed or classified according to the schemes to be tested. Questions are obtained and are translated into the correct terms for searching in each scheme. The retrieved documents are separated into relevant and non-relevant documents, as are the non-retrieved documents. Recall and precision are then calculated by the equations,

$$\text{Recall} = \frac{\text{No. of relevant documents retrieved}}{\text{Total no. of relevant documents in collection}} \times 100\%$$

$$\text{Precision} = \frac{\text{No. of relevant documents retrieved}}{\text{Total no. of documents retrieved}} \times 100\%$$

Deciding whether a document is relevant or not is one of the most difficult problems in such a test. It can be assessed by independent judges, by the people who submitted the questions or by the experimenters and there has been considerable discussion on what is the best approach (4, 5, 6).

The overall conclusions of the many tests so far carried out are surprising. The actual index language or classification used has very little effect on the retrieval performance of a system as measured by recall and precision. Devices to indicate relations, such as links and roles, do not seem to produce a helpful effect on performance. All this seems a very small result for so much effort; it may be that the concepts of recall and precision are too crude, though on the face of it they seem reasonable, or that the tests so far carried out have not really tested the systems. The whole question of where one should go from here has been well reviewed recently by *Robertson* (7).

Up till now, tests have concentrated on two main types of document, i.e. journal articles and reports. Despite the importance of patents as sources of informations, no tests on retrieval performance have been carried out on the specialist patent classification and indexing systems so widely used by industry. We therefore decided to carry out a small-scale test on five such patent classification and indexing schemes in two subject fields, i.e. inorganic chemistry and biochemistry.

2. Patent classification schemes

Patent classifications are generally produced to aid official patent examiners with the legal requirements of their patentability searches. The scope is laid down in the patent laws of each country, and varies considerably. For example, at the time that the patents used in this test were granted, the UK patent search covered only the last fifty years of British patents in an attempt to find a prior disclosure of the invention, or any features of it. In contrast, the US patent examiner has to search through the world's scientific literature, and has to cite prior literature relevant to a patent. The patent classifications can be made to concentrate on the technical rather than legal aspects of patents. The classifications will reflect these different policies.

3. The UK patent classification (8)

The classification is revised every 50.000 published specifications, and this occurs about every eighteen months. A separate 'Reference Index to the Classification Key', which is the index to the classified schedules is revised less frequently and was last revised in 1967.

The UK classification consists of eight sections, divided into forty divisions. The divisions are in turn divided into over four hundred headings, under which are about seventy thousand classifying or indexing terms. Each term is given a number and/or letter code. An example of the classification is shown in Figure 1.

Figure 1: Section of UK Classification

Section C	
Division C3	Macromolecular compounds
Heading C3H	Proteins, etc.
C3HK1	Protein and/or enzyme compositions comprising enzyme-enzyme, protein-protein, protein-enzyme, enzyme-amino acid and protein-amino acid mixtures
C3HK4	Other enzyme compositions
C3HK2	protein compositions characterised by the presence of a functional additive, or a combination of such additives.

Patents are classified by patent examiners to help them in their patentability searches. They therefore tend

to classify features which are novel, or which could anticipate a future claim. Features which are not new but which could be of value in qualifying future inventions may also be indexed. There is also a tendency not to classify by use, as it is difficult to foresee all the possible uses of an invention.
Classifying terms will be taken from as many schedules as the examiner feels are necessary to classify the patent fully.

Listings of all patents classified by particular terms are produced. This was originally done using punched cards but has been computerised since 1970. Some headings have been designed especially for computer searching, so that Boolean logic can be used.

Patent searches are carried out by the public as well as by patent examiners. Such searches can be to determine whether it is worth patenting an invention, or whether a patent is valid. Searches can also be performed for scientific information.

4. The US Patent Classification (9)

The US Patent classification started during the last century and consists of three hundred classes divided into approximately 78,400 official subclasses and 17,000 unofficial subclasses. It grows at about 2,500 official subclasses per year. It is designed to provide for patentability searches by patent examiners. An example of the classification is shown in Figure 2.

Figure 2: Section of US Classification

Class	260	Chemistry, Carbon compounds
	/112R	Proteins and Reaction Products
	/112.5R	Peptides of known chemical structure
	/112.5T	Thyrocalcitones
	/112.5LH	Lutenising Hormone
	/I12.5TH	Thyrotrophic Hornnone
	/112.7	Insulin

Patents are classified into one original classification and a few cross-reference classifications. Reclassification takes place whenever the classification is updated, and this is a continual process. Patent examiners are free to produce unofficial subclasses as often as they appear necessary, and many of these will be incorporated into the official classification by classifiers when they produce a revised classification. Digests are also produced when needed and these are groups of documents of interest to several areas and not any one subclass. These, and unofficial subclasses often cover new technology, especially if it appears in a previously sparsely used class.

The classification schedules are published as loose sheets, which makes updating easy. Sheets of detailed definitions are also available. The index to the classification was last revised in 1972 and has become slightly out of date since.

The system is almost entirely manual. Experiments with automated retrieval have been carried out in some areas.

5. The International Patent Classification (10)

In 1952, the *Council of Europe* made the first attempts to produce an international patent classification. This foundered because it attempted to produce a totally new classification. The International Patent Classification (IPC) developed from a later attempt which used the German classification as its basis. It was finally published in 1968.

The structure of the IPC is similar to that of the UK classification. It consists of eight sections, about six hundred subclasses and 46,000 groups and subgroups. This is illustrated in Figure 3.

Figure 3: Section of IPC Classification

Section	C	Chemistry and Metallurgy
Class	C07	Organic chemistry
Subclass	C07G	Compounds of unknown constitution
Group	C07G 7/00	Proteins; Albumens; Nucleoproteins; Degradation products of proteins.
Subgroups	C07G 7/02	Enzymes
	7/022	Extraction from plants
	7/024	Extraction from malt
	7/04	Metal compounds

The IPC aims to classify any invention, as far as possible, as a whole, and not by separate classification of its constituent parts. The classification can include use if this is an important feature of the invention. Details of general interest in the invention can also be classified by using 'information units', and patents should be classified in several places if they seem appropriate.

The classification was produced by a committee, and this committee has to approve all amendments to the schedules. These can be suggested by all participating Patent Offices. A revised edition, considerably extended in some schedules, was published in 1974, and new editions should be published every five years. The classification was published as looseleaf sheets for the first edition, and as bound volumes for the second. An alphabetical index is published separately.

6. Derwent's indexing languages

Derwent Publications Ltd. specialise in patent information. Their most important product is the *Central Patents Index* (CPI). This covers all the chemical patents issued by 24 major patent-issuing countries. A large number of searching tools are provided for use with CPI and two in particular relate so subject-matter searching, i.e. the manual code and punch code.

The manual code subdivides all chemical subject-matter into 12 major sections (labelled A to M) which are then further subdivided into many classes. An example is given in Figure 4.

Figure 4: Section of Derwent Manual Code

D	Food, disinfectants, detergents
D4	Treating water and sewage
D4A	Treating water
D4A1	Purifying water
D4A1G	Chemical purification of water

The code is intended to be used for simple subject-matter searches or for selective dissemination of information.

The punch code is intended for use on a standard body-punched card with 960 punch positions. The code emphasises chemical structure and activity. It is intended for use with punch-card sorters, or for computerised searching. It is particularly strong on organic chemistry. The code is a fragment code, i.e. punch positions are assigned to significant chemical fragments and are used if that structural component exists in a compound.

7. Previous studies on patent classification schemes

Very few studies have been published on patent classification schemes. *Kento* (11) compared four access methods for searching US patents on the production of vitamin C. One of the four methods used, was the US classification. Only the abstract of this paper was seen (the paper is in Japanese) and from this it is not clear how the four methods were compared or why the author concluded that he preferred using "Chemical Abstracts" (the fourth method of access).

Problems of broad subject searching using the US classification are idscussed by *Murinson* et al (12). They found that 60% of the documents were concentrated in the basic classes, with the remaining documents distributed among a few dozen related classes. If the cross classification were taken into account, then some 30% of the relevant information was lost as a result of this scattering. The authors did not use any recall or precision measurements, but instead showed that their results followed the Bradford-Zipf law. A similar result was found for the IPC when *Murinson* et al (13) also used the alphabetical subject index for broad subject searches. This index would be different from the English index to the IPC, since it would be in Russian, produced independently in Russia. Both these experiments are concerned only with broad subject searches, and no recall and precision measurements are given for any of the searches.

The performance of a detailed subject retrieval system operating with actual user requests, has been analysed by *Shenderov* (14) using patent classifications as an example. Some of the problems encountered in this analysis are discussed, including the problem of working out quantitative measures for assessing recall.

Some work has been carried out on the evaluation of systems, but this was on specific patent index files, rather than on the actual patent classification system. For example, *King* and *Isakov* (15) evaluated the glass technology co-ordinate index file developed by the US and West German Patent Offices. The authors carried out searches, and worked out precision and recall values for each search. These values were averaged and analysed, when it was found that 20% of the retrieval failures were due to indexing errors, whilst 50% were due to overspecific search requests.

An evaluation of the direct coding method in the Derwent FARMDOC system was carried out by *Urhankova* et al (16). They found that in a search, non-relevant information was obtained in 46% of the cases. How the analysis and evaluation of the system was under taken was not stated in the abstract seen (the original paper is in Czechoslovak).

8. Methodology

100 British biochemistry patents and 100 British inorganic chemistry patents which were published in 1976 were selected for study. The selection method has been described in an earlier paper (1). The patents were then classified into the five schemes mentioned above. The British classification was copied directly from the patent specification, as was the IPC. In order to obtain a US classification, US equivalents to the British patents were identified using Derwent's *World Patent Index*. Where US equivalents existed, the patents were looked up and their US classification copied down. US equivalents were found for 51 of the biochemistry and for 52 of the inorganic chemistry patents. The remaining patents were classified using the US classification by the experimenters.

Derwent manual codes and punch codes were assigned by us using Derwent manuals (17, 18) supplied by Derwent Publications Ltd. One of us (C.O.) was already experienced in these codes, as well as in the three classifications.

Questions were then devised by us; 22 such questions were set for the inorganic chemistry set and 15 questions for the biochemistry set. These searches were translated into the five calssifications and the search was then carried out. Details of the retrieved patents were noted.

The entire set of patents was then checked to assess the relevance of retrieved items and of non-retrieved items. Recall and precision values were then calculated. No attempt was made to frame questions so suit patents known to be present and to be relevant, but some questions were deliberately framed to test particular features of the classification schemes, e. g. their ability to retrieve highly specific items or very general items.

9. Results – Biochemistry Sample

Appendix I shows the results for the 15 questions set, and Table 1 summarises these results. The Table demonstrates that the mean recall and precision values for the five schemes are similar, despite the fact that the five schemes differ in their aims and origins and that the two Derwent schemes are more orientated to organic chemistry than to biochemistry. It is also despite the variation in questions from the very general to the highly specific, and that the numbers of patents considered relevant to a query varied from one to 12.

Table 1: Summary of Biochemistry Results

Classification	Mean Recall	Standard Deviation	Mean Precision	Standard Deviation
UK	84.7	19.7	57.9	34.2
US	77.1	26.9	73.5	34.7
IPC	80.7	18.6	68.5	37.0
Derwent Manual	92.1	12.5	73.5	31.4
Derwent Punch	93.9	9.1	58.2	39.8

Examination of Table 1 demonstrates that because of the large standard deviations, the differences between the precision figures for the five schemes cannot be considered to be significant. However, the same cannot be said of the recall figures. The mean recall figures for the two Derwent codes are significantly higher than the three classification schemes, and have smaller standard deviations. The similarity in results could also be due to the questions used and the search strategy adopted. After some searches, certain subjects were chosen to see how well the classifications could cope with them. They were either of particularly broad scope, of narrow scope, or on a diffuse subject but all were thought to be possible subjects for real searches. The number of classification terms varied in each search. For some of the specific ones, a single search term was found exactly to match the question and this was also true for some of the wider searches. Other subjects seemed to have no appropriate search terms, so a variety of possible ones were used. Sometimes the search terms seemed reasonable, but only some of them retrieved relevant patents. Few of the questions covered structural aspects, and this meant that the full power of the chemical structure coding in the Derwent Punch Code could not be investigated. When a question could include chemical structures, a problem was encountered in that analogues could easily be missed if the structure was specified in great detail. If less detail was used, more non-relevant structures were retrieved. It suggests that this coding may be useful for small structural elements but not for more extended ones. Generally only a few punch codes were used in searching. These two features might be thought to lead to a low performance, but this clearly did not occur.

Differences between the classifications can be seen for individual questions and these are due to the presence or absence of terms from the schedules. Precision and ease of searching could possibly be improved by including more specific or general terms at appropriate points in the schedules. Searching would also be helped if there was always a clear indication of the scope of each term. This is available in the US classification in great detail, but less so for the other classifications. Relevant patents were missed because all of the concepts in them had not been indexed. The features were often important to the invention, but not novel, so there is a clash between the patent examiner and information scientists' needs. The use of an invention is often unclassified. Patent examiners clearly have a case for not classifying use, but information will often be needed on devices for solving a problem through any means. The facility to use information units in the IPC could accommodate this. Classification for future patentability searches would be in the invention unit and general informative material in the information unit. However, few countries use information units, and even those which do, often do not assign them to many patents. For example, the British Patent Office only assigned information units to 19 of a hundred random British patents. (BP 1437501 to BP 1437600). An increased use of information units might solve the problem of the different needs of searches through patent literature. They might even be extended to other classifications.

In addition to making comparisons between the classification schemes, the data in Appendix I can also be used to test whether or not there is an inverse relationship between recall and precision. The questions posed varied from the general to specific, and would therefore be expected to give some results with high recall and low precision, and some with low recall and high precision. Study of the many graphs produced by workers in this field indicated that frequently curves are drawn through a scattered set of points on a recall-precision graph and these curves could be disputed. Instead we carried out a Spearman rank correlation coefficient test (19) on our set of results. If a set of recall and precision figures were to follow an inverse relationship law, one might achieve a set of figures such as:

Recall	Precision
100	3
80	20
60	30
30	80

The Spearman rank order correlation coefficient for these two sets of data is -1, i.e. a fully inverse relationship. We fell this coefficient provides a better measure of whether an inverse relationship exists between recall and precision figures than a graph would.

The results of our calculations are shown in Table 2 below. They show that if anything, recall is correlated positively with precision.

Table 2: Correlation Coefficient between Recall and Precision in Biochemistry Set

Classification	Rank Order Correlation
UK	+ 0.20
US	- 0.20
IPC	+ 0.30
Derwent Manual	+ 0.25
Derwent Punch	+ 0.41

A significant inverse relationship requires the rank order correlation to be -0.412 or less.

10. Results – Inorganic Chemistry Sample

Appendix II gives the results for the 22 questions set, and Table 3 summarises these results. The Table shows that once again the Derwent codes give substantially higher recall than the classification schemes, but this time the difference may not be as significant because of the large standard deviations involved. The Derwent codes also are more precise than the classifications, but again the differences may not be significant. There can be no questions, however, that for the inorganic chemistry searches, the IPC performs worse than other schemes.

Table 3: Summary of Inorganic Chemistry Results

Classification	Mean Recall	Standard Deviation	Mean Precision	Standard Deviation
UK	69.7	39.3	56.9	40.5
US	69.5	39.4	53.8	36.8
IPC	49.8	40.1	48.5	39.8
Derwent Manual	80.1	32.1	67.4	35.0
Derwent Punch	88.0	25.9	71.8	31.8

We plotted some recall/precision graphs for the five schemes tested, but these graphs did not seem to show any significant trend. We therefore carried out another Spearman correlation coefficient test on the data, and the results of the test are given in Table 5. Once again, they demonstrate a positive correlation between recall and precision, though this time at a significant level.

A significant inverse relationship requires the rank order correlation is to be -0.360 or less. A significant positive correlation requires a rank order correlation of +.360 or more.

a Significant at 10% level c) Significant at 0.1% level
b Significant at 1% level d) Significant at 2% level

Table 4: Summary of All Results

Classification	Mean Recall	Mean Precision
UK	77.2	57.4
US	73.3	63.7
IPC	65.3	58.5
Derwent Manual	86.1	70.5
Derwent Punch	91.0	65.0

Table 5: Correlation Coefficient Between Recall and Precision in Inorganic Chemistry Set

Classification	Rank Order Correlation
UK	+ 0.36[a]
US	+ 0.62[b]
IPC	+ 0.71[c]
Derwent Manual	+ 0.53[d]
Derwent Punch	+ 0.15

Analysis of the results of the individual classifications shows some interesting differences. For example, let us consider the IPC. When a precise place for a metal compound can be specified, questions on its preparation would be expected to have good recall and precision values. As expected, a compound specifying the amount of water of hydration allowed to be present (Question 14) is too detailed for the system.

The IPC does not always classify the metal part of a compound; when this occurred, the precision figure was low (Questions 14&15). In some cases, the more common metal compounds are specified under the metal (e.g. there is a code for halides of Na/K), whilst in others, there is only one general heading for its compounds (e.g. there is only one code for all gold compounds, C01G 7/00 "compounds of gold"). Even under some of the expanded metal compound headings there is not a place for all possible compounds. To improve the precision, more of the metal compounds could be specified, at least for some of the more common ones.

Some aspects of treating a compound are only classified for certain compounds, e.g. "stabilisation of the y form of sulphur trioxide" is C01B 17/70, but there is no code for "stabilisation of sodium perborate" (seen in the low precision figure obtained in Question 15). Where there is a single code for the treatment of a compound, then the precision (and recall) values are high (e.g. Question 16).

How well the use of a metal compound is classified in the IPC varies. Antiperspirants are moderately well covered (Question 5).

Turning now to the UK Classification, it was noticeable that the examiners in the British Patent Office only classify the inventive part and not everything disclosed in the patent. This is the main reason for recall values below 100% being obtained. The discussion will mainly be concerned with suggestions on how to improve the precision of the searches.

The UK Classification has a code specific to an organic extracting or adsorbing agent, and this term can be combined with the required metal giving the 100% precision figure in Question 2. The electrolytic extraction of metals can also be fairly precisely defined, which accounts for the results of Question 1. The UK Classification will deal with the leaching of a specific metal (Question 3), since the code of the relevant metal can be combined with the process of leaching.

The UK Classification cannot cope with the separation of isotopes (Question 4). A heading specific to isotopes needs to be introduced. This could be combined with the metal of which it is the isotope. Certainly, fewer irrelevant patents would be retrieved in a search on isotopes if a separate code for isotopes was incorporated.

When the combined terms form the required compound, then reasonable recall and precision figures are obtained. In all the 22 search questions, the UK Classification was able to classify precisely the metal part of the compound, unlike, for example, the IPC. Also, in the UK Classification there are no rules on classifying a compound in the last appropriate place, which was necessary in the IPC. This makes the system easier to use.

The UK Classification does not distinguish between the more general aluminosilicate and a zeolite (Question 9), which led to poor precision. This could be improved by introducing a code for zeolites.

As expected, water of hydration is too specific for the classification system (Question 14). Neither the concept of stabilisation nor that of purification (Questions 15 and 16) is catered for in the UK Classification, unlike the IPC. The concepts are indexed under the metal compound concerned, so precision can be low. The high precision for Question 16, occurred because there were no other patents on sodium choloride in the sample. Whether terms need to be introduced or not for these three concepts is debatable since so few patents on these subjects are patented each year (in this sample of 100 there was only one patent on each topic).

On advantage with the UK Classification is that a compound for a particular use or product is always classified under that use or product, as well as under the compound. Therefore general questions where the product is not defined, can often be answered, but the precision may be low.

Using the UK Classification would be made easier by expanding its index. At the moment, the index only refers to the relevant heading, requiring examination of the complete section to find the one relevant term. If several headings need to be searched, then this can be time consuming. Indexing the actual terms appearing in the classification system would considerably reduce the time involved.

Turning to the US Classification, it was found that some of the comments made for the IPC system also apply to the US Classification since they have some aspects in common, e.g. a hierarchical nature. In some of the compounds, for example, the metal present cannot be classified, as in Question 14, giving the low precision.

Separation of metals by extraction (Questions 1 and 2) can be classified, but only by vertical groups (in the Periodic Table), instead of metals individually. This is seen in Question 2, when the loss of precision is due to a patent on silver and gild extraction being retrieved, which with copper, belong to the Group IB metals.

The US Classification only has a special place for radioactive isotopes or radioactive metals of an atomic number above 84. Therefore isotopes of metals of atomic number below 84 would be classified, like the UK Classification, under the metal concerned. This is why the precision is low in Question 4. This classification of isotopes could be improved.

In the preparation of metal compounds, the US Classification can classify a tri- or tetraphosphate, unlike the UK Classification (Question 13), but the metal part cannot be specified. In Question 12, there is no code in the scheme for a vanadate, although there are codes for a titanate and chromate. The vanadate has to be classified under a more general heading. In this case, one can specify the metal present. Since it is a general heading (an alkali and plural metal, oxygen containing compound) the precision is low (25%). To increase the precision, codes for individual compounds would have to be introduced, which would greatly expand the classification system.

Another problem with the preparation of compounds is also seen in Question 12. BP 1431425 deals with preparing alkali vanadates from slag, which is classified under "treating slags". Therefore a search under the vanadate would not retrieve it. Since the question is asking how the vanadate is made, one cannot carry out a search which includes what they are made from!

One problem with classifying a compound under its use, process or product is that for a good precision, the compound needs to be specified under the subject. One cannot combine the product with the compound code, as one can in the UK Classification, since the US Classification is hierarchical. This is seen, for example, in Question 18 when very few transition metal compounds are specified under pigments. Since there is also no particular code for colour, the very low precision (12.5%) was obtained.

The US Classification deals with apparatus in more detail than any of the other four classification schemes. Its code for leaching and extracting apparatus is treated in more detail than the Derwent systems, but all three classification systems had 100% precision and recall values in Question 20.

The index to the US Classification could be improved, and brought up-to-date, since the latest edition is 1972, and many alterations have been made.

The Derwent Manual Code (D.M.C.) is a broader, less detailed and shorter classification scheme than the previous three discussed. This was expected to be reflected in lower precision and higher fallout figures, but this was not the case.

Like the IPC, it classifies isotopes well (Question 4). The process of separation is classified, but like the IPC, does not specify the metal to be extracted (Questions 1 and 2).

Generally, the headings for compounds are broad, covering more than one compound, which can give a correspondingly low precision value for a search. Headings under section E (Chemdoc) can be combined to denote the metal and the anion present in a metal compound, e.g. Question 14. But the codes are still too general for this question (they only state that an alkali and boron are present) and hence the low precision (33.3%) that was obtained. In Question 9, a zeolite is only classified as alumina and silica, the precision and fallout figures reflecting this. The precision could only be improved by introducing codes for specific compounds. This would lengthen the system contrary to the idea of a short, broad classification system. Some precision figures are high, for example, Question 10 (copper carbonate precipitation) since there were no other patents on the subject in the sample. Purification can be specified by combining the process heading for purification, with the required compound (Question 16), but there is no heading for stabilisation (Question 15). A code for this could be added to the classification system.

A compound containing two or more metals is classified in only one place, but according to the rules, the generic codes are also searched. This decreases the precision (Question 12) and did not in fact, increase the recall in this case.

The index to the D.M.C. is better than those of the previous three classification systems, e.g. antiperspirants (Question 5) is indexed. The processes in which catalysts are used are fully indexed (Questions 6 and 7), but the precision is low as the headings often include compositions other than the required one, e.g. under H4-F, Catalysts, there are two codes, one for the composition/preparation of catalysts, and the other an unclassified one. The precision could be improved by introducing more codes.

The results support the idea that broad classes will not give too many retrieved patents to search through for relevant ones.

Although the Derwent Punch Code (D.P.C.) is detailed for metal compounds, these cannot always be precisely classified, e.g. sodium triphosphate (Question 13). Zeolite could be classified precisely, including its property of ion-exchange (e.g. Questions 7 and 9).

In this system a compound is classified under both its composition and its use, process or product. A search involving the preparation of a compound will also retrieve irrelevant patents on its use. This explains some of the poor precision values obtained, e.g. in Questions 9 and 13. It also retrieves the relevant patent under Question 15, unlike the D.M.C. and US Classification. It is the only scheme which includes a code for stabilisers.

When classifying the use of a compound in the punch code, the composition can usually be more precisely classified than in the D.M.C. For example, in Questions 6 and 7, the required catalyst can be specified, under each process it is used in, unlike the D.M.C. Similarly, for pigments and fillers (Questions 18 and 22, respectively), the wanted compounds could be specified.

The D.P.C. is more concerned with classifying a compound than with its uses. The uses tend to be under broad headings, e.g. Question 11, where cements include cement additives and refractory binders.

Although the D.P.C. can cover processes (e.g. purification in Question 16) under the one code there are often several related processes, thus decreasing the precision. This is a result of the limited number of punch positions available on the punch card.

The form of a compound can be specified, but again one code can cover several similar forms. For exemple in Question 17, a single crystal is coded at a position which also includes powders, grains and ground material. This explains the low precision value (20%).

The D.P.C. is better than the UK and the US Classification in dealing with isotopes (Question 4).

There is very little detail for apparatus in the D.P.C., but surprisingly Questions 19 and 20 both had good performance results, because there were very few patents covering apparatus in the 100 patent sample used.

The index to the D.P.C. could be improved. Some terms in the classification system are not indexed, e.g. 56/2 polymerisation catalyst (used in Questions 6 and 7). This means that one has to check through each section every time, but this does not take very long since the sections are short.

The system with the best overall performance is the D.P.C. It is the shortest of the five classification systems, which is inevitable since it is restricted to 960 punch positions. The D.P.C. is a system where terms are combined to form the wanted concept. The UK Classification is also partly faceted, and we therefore expected it to have the next best performance. Instead, the D.M.C. was next, which was rather surprising as it is the second shortest system, with much less detail in its headings than the IPC and national classifications.

11. Misprints and misclassifications

The reason for this section is to comment on some classifications, which through they may be completely justified, from the point of view of a patent examiner, seem strange when the search is for scientific or technical information. These classifications fall into two groups: the classification of patents divided out of a single application, and classifications assigned, or not assigned in general.

The first group became apparent with BP 1346181 to BP 1436184, all of which are divided out of the same application and cover different features of one invention. These are all differently classified and will not all be found using the expected search terms. Similarly BP 1425511 to 1425513 all come from the same application. BP 1425511 concerns a releasing factor and is classified as a protein and medicinal compound, while the other two patents are only classified as proteins because they only concern intermediates for the synthesis of the releasing factor. BP 1447245 and BP 1447246 are a similar case. BP 1447245 is classified as producing an enzyme complex and BP 1447246 as a saccharide although they both concern a method of isomerising dextrose.

From the information point of view, all the patents divided out of an application are relevant and should thus be retrieved. Classifying them in different places makes this difficult. If the specifications can be examined in numerical order, they may be found by glancing at adjacent patents, or at the front of the patent to see if it was divided out. It must be hoped that at least one of the patents will be classified under an expected heading.

Misprints also cause difficulties in patent specifications. A few of the more obvious ones we noted are given below:

BP	1426643	cource for source p.1, column 2, line 70
BP	1429352	CIA S4 92 for S492
BP	1429803	lcaimed for claimed p.5, line 4
BP	1430023	413/126 for 423/126
(US 3897543)		
BP	1433765	In claim 1, half of each line, 90 and 91, on p.3, have been inverted.
BP	1438193	bing for being in the text oxid for oxide in the claims
BP	1438272	solutinon for solution
BP	1438401	C3N 12 missing
BP	1438615	C016 23/00 for B01G 23/00 in the IPC (British version)
BP	1439113	CuO for CaO in the claims
BP	1442617	C01R 31/34 for C01B 31/34 in the IPC (US version)
(US 3976749)		
BP	1448208	B01D 17/66 for C01B 17/66 in the IPC (US version)
(US 3923960)		

According to *Hyams* (20) printing errors in patent office publications are far more numerous than in other types of official or legal documents. Simple spelling misprints are not very serious, but other misprints can have more far-reaching consequences. Classification codes, if wrongly printed, will lead either to the non-retrieval of patents, or to irrelevant patents being retrieved. Some misprints are obvious and would automatically guarantee a closer examination, e.g. BP 1438615. An IPC subclass always consists of a letter, followed by two numerals and then another letter. A subclass printed as C016 would obviously be wrong. In BP 1429352, the space between the terms S4 92 would be noticed and corrected as the other S terms on the patent do not have the space.

Other misprints are likely to pass unnoticed, unless the classification scheme is well-known to the person. For example, there is no such subclass as C01R in the IPC (BP 1442617), and no subgroup B01D 17/66. B01D only extends to B01D 17/10. In the US Classification class 413 has not yet been defined (BP 1430023).

In BP 1438401, the term C2N 12 is missing, as BP 1438402 is on the same topic as this first patent, both dealing with plastic compositions.

Among other printing errors which Hyams commented on, we found passage inversions (BP 1433765) and wrong formulae (BP 1439113).

The US Classification of BP 1430421 (US 3867310) has classified the zeolite as a catalyst. The catalyst is specifically a hydrocarbon cracking catalyst. Normally, if an invention is for a particular use, then it is classified under that use. Why this patent was not classified under this use is not known.

BP 1431508 (US 3914373) is about the separation of isotopes of the same element. In the IPC, B01D 59/24 precisely covers this heading. The US have not classified the patent under this subgroup, but under compounds of the rare earth metals and other metal compounds.

BP 1436524 has classified the zeolite, in the UK Classification, as C1A D41 G12 G4 G4D41. From the rules of the UK Classification it should be C1A D41 G12 G12D41 G4, since it is a double salt of alumina and silicate.

The zeolite in BP 1447102 (US 3929669) has been classified by the US in the IPC scheme as a catalyst. The zeolite claimed, in fact has decreased catalytic activity. This misclassification has probably occurred through the use of the US Classification to the IPC concordance (21). In the US Classification 252/455Z stands for a catalyst or a solid absorbent. The zeolite claimed in this patent has increased absorptive capacity and so is classified under 252/455Z. Use of the concordance has probably only given the catalytic code for the IPC, and this was written down without further checking.

12. Discussion

The results from this study, summarised in Table 4, demonstrate that the Derwent punch code and manual code perform better than national classification systems for the retrieval of patent information in the fields of biochemistry and inorganic chemistry, despite the fact that these two areas of chemistry are not usually regarded as the "strongest" parts of these Derwent codes. However, it should be borne in mind that the sample size (100) was considerably smaller than the numbers of patents that a searcher would be required to search through if doing a full-scale search through, e.g. British patents or a complete run of Derwent records. Thus the precisions recorded in these tests could cause considerable annoyance on a major search.

We have also found what appears to be a weak positive correlation between recall and precision in our results. We attach no great significance to this result, but it does seem to cast some doubt on the traditional truism of an inverse relationship.

We would like to thank Derwent Publications Limited for their assistance in this project.

References

(1) Carpenter, A. M., Jones, M., Oppenheim, C.: Consistency of use of the International Patent Classification.– Studies on Patent Classification Systems I. In: Intern. Classificat. 5 (1978) No.1, p. 30–32, 5 refs.

(2) Robertson, S. E.: The parametric description of retrieval tests. In: J. Doc. 25 (1969) p. 1–27.

(3) Farradane, J.: The evaluation of infonnation retrieval systems. In: J. Doc. 30 (1974) p. 195–209.

(4) Cooper, W. S.: A definition of relevance for information retrieval. In: Inform. Storage & Retrieval 7 (1971) p.19–37.

(5) Cuadra, C. A., and Katter, R. V.: Opening the black box of relevance. In: J. Doc. 23 (1967) p. 291–303.

(6) Rees, A. M.: The relevance of relevance to the testing and evaluation of document retrieval systems. In: Aslib Proc. 8 (1966), p. 316–324.

(7) Robertson, S. E.: Theories and models in information retrieval. In: J. Doc. 33 (1977) p. 126–148.

(8) Tarnovsky, V.: British patent classification and subject searching. In: F. Liebesny, "Mainly on Patents". London: Butterworths 1972.

(9) Glickert, P.: Patent Office classification: its whats and whys. In: J. Amer. Soc. Inform. Sci. 25 (1974) p. 308–311.

(10) WIPO: International Patent Classification. London: Morgan–Grampian 1974.

(11) Kento, I.: Comparison of four methods for the search of American chemical patents. In: Proceedings of the Fourth National Convention for the study of information and documentation (1967) p. 87–91.

(12) Murinson, E. A. et al: Analysis of information scattering in the US patent file and problems of broad subject searching. In: Naučno-techn. inform. Ser. 1 (1976) No.4, p. 18–22.

(13) Murinson, E. A. et al: Use of the IPC alphabetical subject index in broad patent searches. Naučnye i techničeskie biblioteki SSSR (1975) No. 10 (142), p. 19–22.

(14) Shenderow, V. Z.: Some questions of subject acquisition and detailed subject retrieval of patent information. In: Naučno tedhničeskaja inform. Ser. 1 (1972) No.8, p.12–15.

(15) King, D. W., Isakov, P.: Preliminary evaluation of the glass technology coordination index file. In: Proceedings, 7th Annual Meeting of ICIREPAT (1967) p. 229–247.

(16) Urhankova, I. et al.: An evaluation of the direct coding method in the Farmdoc system used for drug patent searches. Met. a techn. inform. 12 (1970) No. 2, p. 18–27.

(17) Derwent Publication Ltd.: CPI Manual Code Manual. London: Derwent Publications 1976.

(18) Derwent Publications Ltd.: CPI Punch Code Manual. London: Derwent Publications 1976.

(19) Siegel, S.: Non-parametric statistics for the behavioural sciences. London: McGraw-Hill 1956.

(20) Hyams, M.: Derwent Patent Services – some problems and special features. In: International Symposium on Patent Information and Documentation 1977. Paper 10e.

(21) U.S. Patent Office: Concordance – U.S. Patent Classification to the IPC. Washington: U. S. Patent Office 1971.

Appendix I: Biochemistry Sample Results

Question No.	Classification Scheme Tested									
	UK		US		IPC		Manual		Punch	
	R	P	R	P	R	P	R	P	R	P
1	75	46	100	89	75	100	100	89	100	22
2	100	50	100	4	100	4	100	100	100	100
3	71	36	86	86	71	71	71	17	86	14
4	100	24	100	30	100	31	100	38	100	67
5	91	56	64	32	55	33	82	60	82	25
6	100	100	100	100	100	100	100	50	100	100
7	67	100	67	100	67	100	100	100	100	100
8	33	12	50	100	67	80	83	100	83	14
9	100	100	100	100	100	100	100	100	100	100
10	100	6	100	25	100	3	100	10	100	7
11	100	100	67	100	100	100	100	75	100	75
12	75	30	38	50	50	33	63	83	75	16
13	100	50	17	100	67	80	83	100	83	100
14	92	58	100	86	92	92	100	80	100	100
15	67	100	67	100	67	100	100	100	100	33

R = Recall (%) P = Precision (%)

Appendix II: Inorganic Chemistry Sample Results

Question No.	Classification Scheme Tested									
	UK		US		IPC		Manual		Punch	
	R	P	R	P	R	P	R	P	R	P
1	80	80	70	78	40	80	70	64	90	48
2	100	100	60	75	20	50	60	43	80	100
3	60	100	20	50	60	75	40	67	100	71
4	100	1	100	25	100	100	100	100	100	50
5	75	75	100	100	75	100	75	100	100	80
6	50	100	75	50	50	100	100	29	75	100
7	60	100	100	63	60	100	100	39	40	100
8	25	100	75	75	25	50	50	67	50	100
9	83	29	80	80	83	56	67	67	100	55
10	0	0	100	100	100	100	100	100	100	100
11	0	0	100	67	0	0	100	100	100	67
12	100	67	50	25	50	33	100	33	100	67
13	100	100	100	100	100	33	100	100	100	17
14	100	33	100	33	100	33	100	33	100	33
15	100	33	0	0	100	33	0	0	100	100
16	100	100	100	50	100	100	100	100	100	50
17	0	0	0	0	0	0	100	100	100	20
18	0	0	100	13	0	0	100	100	100	100
19	100	100	0	0	0	0	100	100	100	100
20	100	50	100	100	0	0	100	100	100	100
21	100	50	100	100	33	25	0	0	100	100
22	100	33	0	0	0	0	100	40	0	0

R = Recall (%) P = Precision (%)