ON THE DESIGN OF AN INFORMATION RETRIEVAL SYSTEM FOR PATENT OFFICE NOVELTY SEARCHING

by

Samuel E. Griffiths

A Thesis Submitted to the Faculty of the DEPARTMENT OF SYSTEMS ENGINEERING In Partial Fulfillment of the Requirements For the Degree of MASTER OF SCIENCE In the Graduate College THE UNIVERSITY OF ARIZONA 1962
STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in their judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Samuel E. Griffith

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

James W. Perry 17 April 1962
Professor of Systems Engineering

Date
ACKNOWLEDGMENT

The author wishes to express his sincere thanks and gratitude to Professor James W. Perry for his patience and deep interest in conducting this study, and for his invaluable guidance and assistance during the preparation of the manuscript.

The author also wishes to express his thanks to Dr. A. Wayne Wyman for his constructive advice not only on this thesis, but also in planning a comprehensive graduate program.

Graduate study at the University of Arizona was made possible by the United States Army. For this opportunity the author is indeed grateful.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Scope of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>2.</td>
<td>REVIEW OF METHODS AND TECHNIQUES</td>
<td>12</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>Old-line Classification</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Patent Office Research</td>
<td>15</td>
</tr>
<tr>
<td>2.4</td>
<td>The Australian System</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>General Advances in Information Retrieval Systems</td>
<td>20</td>
</tr>
<tr>
<td>3.</td>
<td>NOTATION FOR ANALYSIS OF INFORMATION RETRIEVAL SYSTEMS</td>
<td>30</td>
</tr>
<tr>
<td>3.1</td>
<td>Necessity for a Well-Defined Notation</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Basic Notation</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>Symbols to Designate Specific Individuals, Multiplicities, Sets and Specific Members of Multiplicities and Sets</td>
<td>32</td>
</tr>
<tr>
<td>3.4</td>
<td>Notation for Relationships Involved in the Use of Symbols, Languages and Formats</td>
<td>34</td>
</tr>
<tr>
<td>3.5</td>
<td>Notation for Component Parts of Composite Entities</td>
<td>37</td>
</tr>
<tr>
<td>Chapter</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Notation for Processes and Operations</td>
<td>38</td>
</tr>
<tr>
<td>3.7</td>
<td>Conclusion</td>
<td>38</td>
</tr>
<tr>
<td>4.</td>
<td>ANALYSIS OF RETRIEVAL SYSTEMS</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>General</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Classification System</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Uniterm System</td>
<td>52</td>
</tr>
<tr>
<td>4.4</td>
<td>Descriptor System</td>
<td>54</td>
</tr>
<tr>
<td>4.5</td>
<td>ASM-WRU System</td>
<td>56</td>
</tr>
<tr>
<td>5.</td>
<td>ANALYSIS OF BENEFITS AND COSTS</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>60</td>
</tr>
<tr>
<td>5.2</td>
<td>Out-Office Benefits and Cost</td>
<td>62</td>
</tr>
<tr>
<td>5.3</td>
<td>In-Office Costs and Benefits</td>
<td>62</td>
</tr>
<tr>
<td>5.3.1</td>
<td>General Statement</td>
<td>62</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Conventional Classification System</td>
<td>65</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Uniterm System</td>
<td>67</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Descriptor System</td>
<td>70</td>
</tr>
<tr>
<td>5.3.5</td>
<td>ASM-WRU System</td>
<td>72</td>
</tr>
<tr>
<td>5.4</td>
<td>Conclusion</td>
<td>77</td>
</tr>
<tr>
<td>6.</td>
<td>SYSTEM DEMONSTRATION</td>
<td>78</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>78</td>
</tr>
<tr>
<td>6.2</td>
<td>Compilation of a Basic List of Terms</td>
<td>81</td>
</tr>
<tr>
<td>6.3</td>
<td>Semantic Factors and Their Development</td>
<td>81</td>
</tr>
<tr>
<td>6.4</td>
<td>Role Indicators</td>
<td>89</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Scope of the Problem

Design techniques for information retrieval systems have encountered an important and challenging problem in the development of a practical operational system for United States Patent Office novelty searches. The practical nature of the Patent Office novelty search problem requires that due attention be accorded both to the capabilities of various methods and also to the effect of systems design on costs. As a first step in this investigation the pertinent literature on information retrieval methods was reviewed. An appraisal of the state of the general development indicated that automation techniques could be applied to serve Patent Office novelty search requirements. Such applications of information techniques was investigated as a problem in systems design for which guidance was provided by the results of current research in the Systems Engineering Department (Air Force Grant No. AF-AFOSR-61-70 and continuation, Titled: Mathematical Formulation and Computer Simulation of Documentation Methods and Systems, Principal Investigator: James W. Perry). The analysis of the Patent Office search problem was conducted through the present classification headings, classification bulletin, and a random sample of patents supplied by the United States Patent Office.
1.2 Background

The American Patent Office was established to promote the progress of science and useful arts. Our national patent laws are based on the Constitution, Article I, Section 8, Clause 8 which reads:

The Congress shall have the power ... to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive rights to their respective writings and discoveries.

So empowered, the Congress of the United States has been promulgating statute law in this field since April 10, 1790.

The history of the patent system may be traced at least as far back as the middle ages and the various forms of monopoly of those times. The word "monopoly" is derived from the Greek language and means "alone to sell", whereas the phrase "letter patent" is derived from Latin and means "open letters". Organization of trade as monopolies was in general vogue during the middle ages. A sovereign or a state often granted a monopoly for one or more purposes, such as the establishment of commerce, trade and industry, or the production of revenue. In the thirteenth and fourteenth centuries, the guilds and trade unions achieved powerful monopolistic positions which enabled them to dominate trade. The trade union and guild monopolies were generally broken in England during the fifteenth century, but the Crown continued to grant monopolies in all fields to favored individuals. As a result of the abuse of this power, the parliament and courts started, early in the seventeenth century, to restrict the granting of monopolies. In 1615 the English courts held that the Crown did not have the general power to grant a monopoly but if a man had made a new discovery,
or introduced a new trade into the realm, the Crown in recompense of
his expenditure may grant him the exclusive rights therein, for a lim-
ited period of time. All other types of monopolies were abolished by
Parliament in 1623 by enactment of the Statute of Monopolies.

The obnoxious monopolies, against which Parliament acted in
1623, had been granted in the form of letters patent. Letters patent
had also been granted for inventions and the Crown was permitted to con-
tinue to do so under the Act of 1623. However, the history of the term
"letters patent" or "patent" sometimes leads to confusing a patent
granted for an invention with the older obnoxious monopoly grant, which
made it possible for some favored person or organization to block the
public from some previously developed activity in the general field of
manufacturing, industry or trade. In commenting on this point, Walker
on Patents states that:

A patent is not a monopoly, it is not a grant in deroga-
tion of some common right, the patentee takes nothing from
the community; he is a great public benefactor, because he
gives to the community his invention for the reward provided
by the statute in the form of Letters Patent granted to him,
which is in effect, a contract between the inventor and the
public.

The exclusive privilege granted to an inventor, is not as generally
understood the exclusive right to make, use or vender the invention of
the patent but, on the contrary, is the right to exclude others from
making, using or vender the patented invention, without the permis-
sion of the patentee. A patent, accordingly, is sometimes regarded
as the right granted to an inventor in exchange for the disclosure of
his invention to the public. It is, indeed, the case that much
The time period during which an inventor may enforce his patent rights is a matter for regulation by statute and is currently set at seventeen years. The same is true of the kind of subject matter that may be patented. Regulations have also been set up that specify procedures for obtaining a patent. Such laws and regulations differ considerably in different countries.

What may be patented in the United States? The United States statutes state:

Any person who has invented or discovered any new or useful art, machine, manufacture or composition of matter, or any new and useful improvement thereof ... not known or used by others in this country, before his invention or discovery thereof, and not patented or described in any printed publication in this or any foreign country, before his invention or discovery thereof or more than one year prior to his application, and not in public use or sale in this country for more than one year prior to his application, unless the same is proved to have been abandoned, may, upon payment of the fees required by law, and other due procedures had, obtain a patent therefor.

Due to the fact that patents are established by statutes, they have been the subject of many judicial actions. As a result of these judicial actions, the meanings of the statutes and of their terms have been very carefully defined. What may be patented has been spelled out in great detail. Some of the more important of these actions have lead to the following definitions and restrictions:

1. No discovery, which does not in effect amount to a contrivance or production of something which did not exist before, or in other words to an invention, will entitle the discoverer to a patent.
2. Art is, in the sense of the patent law, synonymous with a process or a method when used to produce a useful result.

3. The term "machine" within the meaning of the patent law, embraces not only machines as commonly understood, but various mechanical elements and combinations.

4. The term "manufacture" has a very comprehensive sense, embracing whatever is made by the art or industry of man, not being a machine, a composition of matter, or a design. But while articles of manufacture may be new in the commercial sense, they are not new in the sense of the patent law and are not patentable as new manufactures, unless it appears in the given case that the production of the new article involved the exercise of invention or discovery beyond what was necessary to construct the apparatus for its manufacture or production.

5. The patentability of an article of manufacture is not affected by the fact that such article can be produced on machines previously in use. But to obtain a patent for a product made from raw materials, it must possess a new or distinctive form, quality or property.

6. The phrase "composition of matter" covers all compositions of two or more substances. It includes all composition articles, whether they may be the result of chemical union, or of mechanical mixture, or whether they be gases, fluids, powders or solids.

7. An improvement is some addition to or change in an existing machine, art, manufacture or composition of matter. An improvement may lie in addition, simplification or variance but must be new and useful.
In line with decisions on what is patentable, there have been decisions on what is not patentable. In general, the following are not patentable:

1. Laws of nature and scientific principles
2. Abstract ideas
3. Mental theories
4. Plans of action
5. Systems of business
6. Functions, results or effects of machines

Why novelty searches? The patent statute of the United States have always provided that statutory subject matter to be patentable must be new and useful. Novelty and utility are therefore two essential requisites of patentability. Hence, it is the responsibility of the Patent Office to examine every application for patent to determine if novelty has been established. The criteria of novelty, as applied to these statutes, are very broad, in fact, broader than the term "new" in the ordinary sense. Such breadth is the result of judicial decisions, and therefore the best method for taking note of its boundaries is to cite some of the more important judicial interpretations.¹

1. Novelty is not negatived by anything not substantially identical with the subject of the patent, even though the function of the prior process or thing was identical with that of the patented matter.

2. Novelty is not negatived by prior devices incapable of functioning beneficially.
3. Novelty is not negatived by prior knowledge and use in foreign countries, unless patented or described in a printed publication.

4. Novelty is negatived by prior knowledge and use in this country, even by a single person.

5. Novelty is not negatived by mere experimental use.

6. Novelty is not negatived by forgotten or lost art.

7. Novelty is not negatived by unappreciated prior use but is negatived by demonstratable prior use.

8. A printed publication is anything which is printed, and, without any injunction of secrecy, is distributed to any part of the public in any country, and such a publication may negative novelty.

9. Disclosure in a prior patent must be full and clear, it can not be vague.

10. Mere crudity of a prior device does not prevent it from being an anticipation.

11. Mere omission of an element which may be supplied by a skillful mechanic, is not fatal to anticipation.

12. Novelty is not negatived by any prior abandoned application for a patent.

13. Novelty of a machine or manufacture is not negatived by any prior unpublished drawings, no matter how completely they may exhibit the patented invention.

14. Novelty is not negatived by antiquity of parts. The fact that all elements are old is not conclusive against patentability, nor does it preclude the possibility of invention in the new combination.
15. Novelty is not negatived by any prior accidental occurrence or production, the character and function of which was not recognized until later than the date of the patented invention, sought to be anticipated thereby.

16. Where an invention has passed away from the memory of the one who made it, and of those who have seen it, and the object itself has disappeared, then the knowledge of such invention is said to be completely lost as if it had never been discovered. If another, by his own efforts, without any knowledge of the previous invention, re-invents the same, he is discovering something which is then new and unknown and is entitled to a patent therefor.

Obviously, these requirements as to novelty and subject matter make it necessary for each application to be reviewed with care by the Patent Office. In particular, Patent Office examiners must conduct a careful and comprehensive survey - a "novelty search" - of previous publications that read on or are closely related to the subject for which the applicant is requesting patent protection. As human knowledge continues to accumulate and, in particular, the number of patents increases, novelty searches have become more difficult. In fact, the United States Patent Office, through work-unit evaluation, has observed that there has been a continuing decrease in the number of applications brought to final action per examiner per year. This number has decreased from 230 final actions per examiner per year in 1900 to less than 95 at present. A management survey of Patent Office operations made some ten years ago revealed that something more than 60% of the time of the examiner was devoted to searching, that is to say,
to identifying which patents or other publications may be of pertinent interest to the individual applications. Patent applications are being filed at the rate of 80,000 per year in our Patent Office. It was pointed out in the Bush Committee Report of December, 1954, that a critical situation had developed. 18

Up to the present, the only major solution has been the hiring of more examiners, but even so delays of from two to four years in issuing patents after the receipt of the application are so frequent that they might be termed normal. All of the above is contributing to the increased cost of operating the United States Patent Office, and it is possible to foresee, in the not too distant future, a situation in which it would no longer be financially feasible to make a thorough novelty search prior to issuing patents. Yet, as we have seen, a novelty search is essential to safeguard patents from becoming obnoxious.

In some countries, e.g., Switzerland, patent applications are made available to interested parties who may then present reasons, especially prior publications, that would argue against granting patent protection for all or some of the subject matter as requested by the applicant. In the United States, the practice has been to keep applications secret until final decision as to patentability has been made. Departure from this practice seems likely to have grave repercussions for our patent system.

Abolishing patents altogether can scarcely be seriously considered. The Dutch Parliament (States-General) in 1869 repealed their patent law, but due to the slowing in technical advances, the Dutch Patent Office was re-established in 1910. There can be little doubt
that closing of our Patent Office would not be an acceptable solution to the novelty search problem.

In addition to novelty searches, which are particularly numerous and extensive, the Patent Office also conducts other types of searches, in connection with interferences and to support judicial decisions in infringement cases. Numerous "state of the art" and "pre-ex" (prior to filing the application) searches are conducted for industry. Such searching service provided by the Patent Office to industry brings out the very important point that the novelty searches of the Patent Office are closely allied to the research and development searches carried on by industry. In reality, patents are for the most part the result of industrial research and development, and contain information of the same general type. Industrial firms usually conduct literature searches when planning and conducting research and development projects. Rising costs of such searches have impelled development of automated information retrieval systems. A number of companies, as well as governmental and other organizations have demonstrated that automation of information searching can provide benefits which are beyond the capability of ordinary indexing and classifying methods. The Patent Office, itself, has become aware of this fact and has had underway for nearly ten years, a research and development program in this field, to develop an automatic information retrieval system to expedite novelty searches and the like. Although some practical system development work has been done in a few restricted fields, no general approach to the problem has yet been set forth. Important basic problems remain unresolved, especially questions of research strategy in
developing generally applicable methods to meet United States Patent Office requirements. Work on automated searching methods, completed and published to date by the Patent Office, has all been in the chemical composition field and doubt has often been expressed whether methods could be developed for advantageously applying automation to novelty searches in the mechanical and electrical fields.
2.1 Introduction

In the preceding chapter attention was directed to the essential role of the novelty search in our patent system. It was noted that at least sixty percent of the examiner’s time is taken up by searching, that is, in identifying which documents and patents are of pertinent interest to various individual applications. Other persons and organizations in making use of scientific and technical literature have encountered similar problems. Recognition of a general bottleneck situation in searching dates back at least fifteen years. During this period traditional methods of classifying and indexing have been studied as to capabilities and limitations. Simultaneously much effort has been directed to developing new methods for selecting information from a composite grouping of documents. The results of such investigations are reviewed in this chapter.

2.2 Old-line Classification

The term “classification” as used here is defined in the classic form as in Webster’s Dictionary: “Act or result of grouping and segregation into classes which have systematic relations.” Such relations are expressed by headings and subheadings which serve as labels to indicate the contents of classes and subclasses organized
in fixed arrays. A more extended use of the term has come into use to include the coding of material for retrieval systems such as punched cards or the results of searches thereon.

Classification, in the more restricted sense of the term, may be regarded as the development of systems for grouping together like items of information in an orderly and useful arrangement of classes and subclasses. The divisions of a classification system should add up to the total subject matter being classified. Each division, in turn, is made up of a number of subdivisions adding up to the entirety of that division, and so on down to the more minute subdivisions as may be necessary or desirable. It has been a precept in the Patent Office that a classification system worthy of its name is organized in such a way that the material is not only broken down into relatively homogeneous groups, but the groups are arranged so that relationships between the groups are apparent upon scanning the schedule of class titles.

It has been generally recognized that there is no uniform proper order for arrangement of classes and subclasses. This is noted in the introduction to the Universal Decimal Classification System:

Many classifications, of knowledge in general and science in particular, have been evolved rather as an intellectual exercise than with any specific aim in view. Numerous attempts have been made to derive so-called logical classification, designed to satisfy the ideal criterion that the order of sequence of the classes shall correspond with the logical order of sequence of the concepts represented. It is assumed that a logical or natural relative order of the various branches of human knowledge and activities exists and can be determined. That this assumption is false and the pursuit of the ideal illusory is demonstrated by the rapidity with which such attempts have lapsed into oblivion. On the other hand, classifications
which have initially been designed to meet a specific practical purpose are nearly always found to have a reasonably logical sequence and to be readily adaptable to other purposes, provided the latter are not too remote in characteristic from the original function for which the classification was designed.

Awareness of these considerations has guided the Patent Office in attempting to develop the classification system so as to provide the greatest measure of usefulness to the greatest number of searches.

The United States Patent Office classification system, currently in use, consists of 311 major classes and in excess of 57,000 subclasses. This system has been built up over the years and is periodically revised to reflect important changes in technology. Since the system reflects both years of experience in classifying patents and also trends in technology at the time of establishing various classes, this system was used as the starting point for this study.

The Patent Office classification system has been worked out so as to provide a place for each item classified. To serve this purpose each of the 311 major classifications includes at least one miscellaneous subclass.

In order to make this classification more precise and hence more effective, the Patent Office publishes Classification Bulletins. These bulletins are issued every time a new class is added or new subclasses created. The Classification Bulletins give a written interpretation of each of the headings contained in the classification system. This is done in order to counteract ambiguity and to provide uniformity in the interpretations.

In addition the Patent Office maintains an extensive
alphabetized index which is cross-referenced to the classification system. This is done to facilitate novelty searches and to further increase the effectiveness of the classification system.

To enhance the usefulness of its classification system even further, the Patent Office strives to set it up so that the classification shall consist of mutually exclusive groupings. In other words, the Patent Office strives to provide one and only one location for each of the items being classified. The present classification system does not achieve this ideal completely and the multiple entry of individual documents in different subclasses is a factor that contributes importantly to such practical difficulties, as costs incurred by searches. Even more important, as a cost factor, is the fact that new patent applications often involve more or less extensive regrouping of concepts. These difficulties and related practical considerations are taken into account in analyzing costs in Chapter 5.

2.3 Patent Office Research

The United States Patent Office has been conducting research and development studies on automation of novelty searches of its own for about twelve years. These studies have involved automation of individual classes, general automation procedures and equipment development.

The first experiment in automation of novelty searches dates back to 1950. This experiment was concerned with medical compositions. It was demonstrated that machine searching could be performed so as to detect patents of pertinent interest. However, the costs of
encoding by the methods of this early experiment were so high that practical operations required further methods development.

In 1951, hearings conducted by a committee headed by Vannevar Bush led to the recommendation that the Patent Office should develop searching procedures employing automation techniques in order to cope with the novelty search problem. Subsequently, studies to implement this recommendation were started and, early in 1956, an Office of Research and Development was established in the Patent Office. Since 1957 approximately ten research and development reports have presented results of studies which include automating of searching in a variety of patent subclasses in the general field of chemistry.

Special coding and searching systems have been developed and operationally exploited for two narrow fields. The first of these in 1957 was a punched card system for steroids, chemically derivatives of cyclopentaphenanthrene. The coding was based on numbering the carbon atoms in the ring system and then indicating atoms other than hydrogen as well as functional groups attached at various ring positions. The system has been slightly modified recently to expand the range of compounds covered. This coding system has functioned very well within the narrow field for which it was designed. The time lag in processing of applications has been considerably reduced. This coding system is restricted, however, to steroids and such a system can scarcely be regarded as a general prototype searching system, even in the field of chemistry.

Organic phosphorus compounds constitute another narrow field of chemistry for which a specially designed coding and searching
system were developed in 1960. The system was designed to operate on the RAMAC computer and employed several of the features that had been found useful in other experiments. It employs the inverted file technique (further discussed in section 2.5), use of nodes to show relationships and atom groups as building blocks. The results of this system have also been very satisfactory to date.

These developments of specialized chemical coding systems were accompanied by development, in cooperation with the Bureau of Census, of the Integrated Logic Accumulating Scanner. This is a row by row scanning device for punched cards which was designed to handle data processing problems in both offices.

To date, eight research and development reports of more general nature have been published by the Patent Office. In addition to general Patent Office problems, much attention has been directed to linguistic problems, notational systems and other general retrieval difficulties. Simon N. Newman, in a series of three reports, described how the development of a retrieval system for all classes of patents might be undertaken. He recognized that for a retrieval system to function effectively in the Patent Office, it must show how terms are interrelated and in particular relationships between generic and specific concepts. The Newman reports also recognized the importance of designating another class of relationships, namely those expressed by phrasing and sentence structure in ordinary language. These same general principles were applied, though in different embodiment, in the ABM-WRU system to be explained subsequently. However, successive research papers issued by the Patent Office have proposed more extensive
ramifications of the coding methods to be developed and the resulting complexity of coding seems certain to involve excessive costs. These later papers, which do not report any actual applications, appear to be primarily of theoretical interest.

All of the above demonstrates the interest of the Patent Office in the development of automated retrieval systems. However, the nature and scope of this work and its very limited operational application indicate the need for further explorations into practical approaches to the novelty search problem. The operation of a large number of independent retrieval systems for different subclasses does not appear to be the most practical approach, as this seems certain to lead to duplications of effort and files and most probably excessive operating costs. The general approach to the problem, as taken by Newman, appears to hold the most promise for an effective total system, provided that operating costs, especially the costs of analyzing and encoding patents, can be held down to practical levels.

2.4 The Australian System

The Australian Patent Office has encountered the same general type of problem in novelty searching as the United States. Although the number of applications filed annually in Australia is approximately ten percent of the number filed in the United States, a start toward automation of novelty searches was made over fifteen years ago. Two examiners, K. B. Peterson and T. K. Hogan, were given time off from their normal duties to derive a code, encode material, and demonstrate practical utility. As a result of this work, the Australian Patent
Office adopted an automatic punched-card system. An immediate complete transfer to the new method was not attempted. Instead, one class at a time was converted to the new system starting with the more active classes. The experience of the Australian Patent Office shows that search times have been reduced by at least fifty percent.

The class chosen for the original work was one pertaining to electronic devices, fluorescent lamps and related devices which totaled more than two thousand patents. The system was designed so as to make use of standard British Tabulating Machine Company equipment, which employs punched cards with eighty columns of twelve positions per column. The codes were worked out so that the punching of a single hole records some one characteristic of subject matter. The characteristics may be complex in character, e.g., the presence of a ring system in an organic compound.

The code format for the cards is: Columns one through ten are reserved for classification number and year. The remainder of each card's eight hundred and forty positions remained available for characteristic coding. Approximately six hundred of the characteristic positions were used in the original plan, leaving space for future code expansion. To maintain resolution and discrimination as required by future searching requirements, it was realized that the code for a given class could not remain static but would have to be modified to cope with material constantly being added to the class.

The system as evolved still makes limited use of conventional classification system in that patents are placed in distinct though general classes. Within a given class mechanized searching selects
those patents of interest to a given application. The basis of such searching is the punched-card recording of the characteristics of each patent in the class. With this system it is desirable to have no more than twenty-five hundred patents per class, as the degree of selectivity attainable with this equipment (punched cards and sorters) is limited.

This system provides eight hundred and forty different characteristics for each class and these may vary, of course, from class to class. If a patent, because of its scope of subject matter, has to be assigned to two classes, it is necessary to make two cards for it. It requires one card for each class since characteristics used in the two classes will be different and consequently the meanings assigned to the respective holes in the punched cards will be different in each class.

This appears to represent a start in a highly promising direction, but if the favorable results attained with a small system are to be realized with a large one, the system must be made considerably more general and versatile. At the same time, the selectivity must be correspondingly increased while avoiding excessive costs in the analyzing and encoding of the subject contents of patents. Even though relatively simple in form, the Australian method does demonstrate the usefulness of automatic practices in Patent Office novelty searches.

2.5 General Advances in Information Retrieval Systems

The importance of information retrieval problems has stimulated a considerable variety of empirical developments whose nature
and scope can be best presented by considering three typical systems in some detail.

As will be evident, each of these systems may be varied as to range of characteristics that may be used for analyzing documents, and they provide considerable flexibility in meeting different requirements. For each of these systems, various types of equipment can be and has been used. The effectiveness of any information retrieval system is primarily determined by the type of code used, since this is what provides and determines the selectivity and effectiveness of the system. Once the code has been established, the selection of the hardware will influence and determine the following: Speed of operation; the number of searches that can be conducted simultaneously; cost of operation. Accordingly, in this investigation attention was centered on selection and development of the code system. Selection of the hardware was regarded as a further separate problem, whose solution may be deferred until the code system has been developed, work loads determined and required operating speeds specified. This approach requires that the code system be developed so that hardware requirements remain within the range of available equipment. In this connection, experience provided by previously developed information retrieval systems furnished the necessary guidance.

The first system to be reviewed, the Uniterm system developed by Mortimer Taube, makes use of individual designators of information called "Uniterms". A uniterm may be a single word, phrase, arabic numeral or any other information representation. Usually uniterms are individual words selected directly from the documents being processed.
These words are selected by an operation analogous to indexing; that is, by choosing key-words which are descriptive of ideas or concepts embodied in the documents.

The uniterms are recorded for subsequent use in performing selections by establishing for each uniterm a separate field, either a separate card or a zone reserved on a continuous record. Within a given field, there are then recorded serial numbers or similar identifications of documents to which this field's uniterm pertain. In other words, this system employs the inverted file technique; that is, that the basic entry in the system is not made under the document heading, but rather under the uniterm heading. A search is performed by selecting the fields of all uniterms which pertain to a given query and then determining which document serial numbers or the like are contained in all these fields. Any document that is referred to in all the fields is selected as pertinent to the particular query.

The uniterm inverted file technique requires that as documents are analyzed their serial numbers or the like shall be properly entered in the various uniterm fields. This requirement may be met in various ways. For example, document serial numbers may be posted on cards or on magnetic tape or disc. In the latter case, provision must be made to record newly entered serial numbers, while avoiding excessive extensive blank areas in the recording medium or disturbing the existing file.

In this system, the documents are analyzed directly in total text form by selection of the key words or phrases that best represent
the subject content of the document. The document is thus characterized by the list of uniterms. It is sometimes the case that no control is imposed on the uniterms and in this way an unlimited number of terms may be introduced into the system. It is to be expected that such a procedure will result in the same idea or concept being referred to by different terms in different documents. Such use of alternate uniterms has been observed to cause difficulty when conducting searching and selection operations. As a consequence, the desirability of working with controlled terminology has come to be widely recognized as an important factor in attaining desired benefits with the uniterm system.

As may be evident, a search can be directed, in principle at least, to any logically defined combination of uniterms. However, practical limitations are encountered with respect to the number of uniterms used in combination to specify a given query. The degree of practical flexibility is, however, much greater than attainable with the headings of conventional classification systems. Such headings provide for much fewer fixed combinations of ideas and concepts than can be readily generated by establishing combinations of uniterms.

The uniterm system, by virtue of its inverted file technique, does not provide simple means for recording either (i) relationships between generic concepts and specific terms or (ii) syntactical type relationships as designated by constructing phrases and sentences in ordinary language.

The second system to be considered, the descriptor method of Calvin Mooers, establishes for a given subject field a limited number
of concepts which usually are neither highly specific nor broadly generic in character. The inverted file technique is not used but, instead, all the descriptors for a given document are recorded as a block, e.g., by punching of a card or as a sequence of entries in a magnetic recording medium. This obviates problems of entering codes for new documents as discussed above in connection with the uniterm system. On the other hand, the scope of a query may be designated by logically defined combinations of descriptors. The selectivity of the system is determined primarily by the choice of descriptors when setting up the system, rather than by restrictions as to the number of terms in combination of descriptors for which searching equipment can conduct selections. This is a direct consequence of the fact that the descriptor system was originally set up to work with simple punched card equipment.

The descriptor system analyzes documents in their total text form and selects important features of the subject matter which may be expressed directly by appropriately selected descriptors. Alternately, the important features of subject matter may first be expressed as key words, phrases, or the like and the latter then converted into descriptors which thereafter characterize the document. In order to maintain consistency within the system, it is often desirable to establish a thesaurus for listing the descriptors and their inclusive range of key words, phrases, and ideas. A basic principle is to keep the number of descriptors to a minimum for the field of knowledge covered. As a consequence the range of terms, i.e., the
vocabulary, of a descriptor system tends to be less extensive than in applications of the uniterm system.

Preparatory to conducting a search, the scope of a given query is expressed as a combination of descriptors. A search is performed by comparing the descriptors for each document with those that typify the query, and any document whose set of descriptors contains all the descriptors of the query is considered to be of pertinent interest.

The third system was developed by James W. Perry and collaborators, under sponsorship of the American Society of Metals at Western Reserve University. This system is much more elaborate than the previous two and requires more detailed description. The basic principle is, however, much the same. The subject contents of documents are analyzed and expressed by various characteristics that are recorded in encoded form. The scope of a given query is expressed as a combination of characteristics. The search-selecting operation consists in detecting those documents whose combinations of characteristics include as a subset the combination of characteristics for the query being searched. When such an inclusion relationship is detected, the document in question is selected as being of pertinent interest. One major difference of this system is the use of multi-level Boolean algebra in contrast to the use of one level by the two previously summarized systems.

As with the other two systems, the total text of each document is reviewed, but the important features of the subject contents are expressed as an abstract which may be written directly in a standardized form of phrasing, preparatory to being coded.
The various terms in the abstract are then encoded using a thesaurus type code dictionary to which attention will be directed next.

Individual words and phrases are not used directly as designating units as in the uniterm or descriptor systems. Rather each term in the standardized abstract is recorded as a code which is unique for each term but which indicates how the term in question is related to those generic concepts which pertain to the term. The generation of such codes is based on a set of generic concepts that are worked out as being both important to the field of knowledge involved and also as being effective for searching and selecting operations. These generic concepts, or semantic factors, and codes generated from them may be illustrated by the code for thermometer whose three semantic factors designate a thermometer as: (i) a device (ii) for measuring (iii) heat. The codes for individual terms are made more meaningful and precise by associating with each semantic factor a further code element that indicates how each semantic factor is related to the encoded term. Both these relationships as well as the semantic factors may be selected and defined as may appear most appropriate for the subject area for which a code system is being developed. In order to provide unique codes for closely related terms, a distinguishing numeral is made a part of such codes. For example, thermometer may be represented by MACH (device), MUSR (measuring), RWHT (heat), 002 and pyrometer may be defined as MACH, MUSR, RWHT, 004. With this system, individual semantic factors or their combinations may be used to specify one or more ranges of terms, that is to say one or more classes of terms when specifying the scope of a query. Such classes of terms may
be regarded as corresponding to the descriptors of the second system previously reviewed. New codes for specific terms may be generated as required when analyzing documents. This system thus combines generic coding with the ability, unambiguously, to record specific terms and to use them as appropriate for expressing the scope of a given query.

The ASM-WRU system also provides for indicating relationships of a syntactical nature as usually indicated by phrases and sentence structure in literary language. Such relationships may be set up in different ways when working out a given code system. In any case, such relationships are similar to those that are revealed by diagramming of sentences in basic English to show nouns, verbs, adjectives, adverbs and relationships among them. However, this code system specifies physical relationships directly by means of code elements known as role indicators. For example, a starting material will be assigned the same role indicator even though this relationship may be indicated by different phrasing and different sentence structure in literary style abstracts. The role indicators are designed to serve as a convenient means for specifying standardized phrasing and standardized sentence structure in the encoded abstracts for machine searching.

The encoding of documents may now be summarized as consisting of the following steps:

1. Analysis of the documents as to the important features of its subject contents. (Establishment of policy decisions as to what features of subject contents are to be regarded as important is an essential feature of developing a specific operational system.)
2. Expression of the important subject contents features of each document as a standardized abstract with the aid of role indicators in accord with rules for standardization of phrasing and of sentence structure.

3. Replacement of each of the terms in the standardized abstracts by a corresponding code for each term. (Such codes consist of semantic factors together with a numeral to ensure a unique code for each term.)

4. Recording the results of such encoding in an appropriate medium, e.g., magnetic tape, for subsequent machine searching.

The last two steps can be readily performed by automation procedures supplemented by human effort to generate codes for newly encountered terms and to resolve questions as to which code may be appropriate for a term, such as "cell", that may refer to widely different concepts. The first two steps, as above outlined, are, as a rule, best performed by people who can bring to the analysis of documents an understanding of their subject contents.

As may be evident, newly acquired documents may be processed at any time and their encoded abstracts recorded, e.g., on magnetic tape, without disturbing those previously recorded. If desired, the encoded abstracts may be grouped as to subject field to reduce the number of such abstracts that must be scanned to service a query in a given subject field.

The encoded abstracts consist of strings of symbols organized at successively higher levels. In the encoded abstracts, the codes
for terms in conjunction with role indicators are used to build up code combinations analogous to phrases in ordinary language. Special rules are established for generating such phrases and for organizing them into higher order structures analogous to sentences and the latter, in turn, into paragraphs or combinations of paragraphs.

A query may be encoded to specify that one or more combinations of semantic factors shall characterize one or more sets of terms, that correspond to descriptors in the Mooers system. Furthermore, the encoding of a query may specify that the terms, as specified by combinations of semantic factors, are required to be associated with certain role indicators. In addition, the phrase, sentence and paragraph organization of abstracts to be selected as pertinent may also be specified. Such syntactical association among terms contributes to selectivity during searching, by making it possible to discriminate between an encoded abstract which contains a combination of terms in one syntactical pattern from another encoded abstract which contains the same combination of terms but in another syntactical pattern. Syntactical association also eases the problem of dealing with compound terms and, in addition, it makes it possible for the context effectively to define such otherwise vague terms as "high", "low", "hard", "soft", "fast", "slow", etc. The use of techniques included in this system provides greater flexibility than either of the others, and by using automation techniques to accomplish much of the encoding, the increased margin of flexibility can be attained without incurring excessive costs.
CHAPTER 3

NOTATION FOR ANALYSIS OF INFORMATION RETRIEVAL SYSTEMS

3.1 Necessity for a Well-Defined Notation

In the preceding chapter, a general description was provided of conventional classification and of more recently developed information retrieval system that appear promising for practical application to meet the novelty search requirement of the Patent Office. Analysis of the novelty search problem and evaluation of previously developed methods are essential steps in designing an information retrieval system to meet Patent Office requirements. Such analysis must be both realistic and precise. This analysis cannot be carried through unless we have available the means for specifying various operations and their results, as well as stating common characteristics and distinguishing features of the information retrieval methods.

This chapter summarizes a general notation for analyzing information retrieval methods and systems. This notation and its applications have been developed during the past eighteen months at the University of Arizona under an Air Force grant.

3.2 Basic Notation

D is reserved for designating individual members of the general class of documents; in other words, a graphic record or some specific portion of a graphic record as may be appropriate in one case or
another. (For the purpose of this investigation, a document is taken to be a patent.)

T is reserved for designating any set of documents and is used when attention is being directed primarily to the set rather than to the member documents.

Q is reserved for designating any query.

S is reserved for designating any symbol, especially a symbol assigned to some D, T, or Q. In general, a given S constitutes an organized structure and may be anything from a single mark to the total text of a document.

E is reserved for designating an element within an S. In discussing how E's are used to generate S's and especially interactions between S's, we will have occasion to characterize various kinds of E's; for example, those that refer to individual physical objects, to operations, to syntactically expressed relationships of an observational nature.

L will be used in various discussions with appropriate subscripts to designate various individual languages. A language is considered to consist of a set of E's together with a set of rules for making use of them to generate S's. A change in either the set of E's or in the set of rules means that the language has undergone a change.

F is reserved for designating a format in which a symbol may be recorded. Alternate modes of recording the same symbol in some one language are frequently encountered in information retrieval. The symbol for a given code, for example, may be expressed in typewritten
form or by punching in a card. The conversion from one format to
another may be required in a given system design.

A is reserved for designating an array, in particular, of T's
and S's. Attention will be directed not only to the structure of clas-
sificatory and related arrays but also to the processes and rules for
setting them up.

P will be reserved for designating a process which will often
be further specified or characterized for clarity.

U is reserved for designating an operation which is comprised
by a process. More precisely and more formally, we will consider a
process P as consisting of a set of operations. As far as input and
output are concerned, one process may provide the same performance as
another even when the two processes consist of different sets of U's.
From both a theoretical and a practical point of view, it is often
the case that a U may be analyzed further into suboperations. Thus,
there is an element of arbitrariness that is controlled by practical
considerations as to the degree of analysis of a given U and conse-
quently in their definition.

R is reserved for designation of a rule which specifies any
one of the following: (i) How to perform some P or U, (ii) how to
test the correctness of performance of a P or U, and (iii) how to de-
tect errors in performance or in results of performance.

3.3 Symbols to Designate Specific Individuals, Multiplicities, Sets
and Specific Members of Multiplicities and Sets

As already noted, individual letters, such as D, Q, etc., have
been reserved to indicate some one item of a certain kind as these may be defined in a given practical situation. Specific individuals are denoted by subscripts. Thus, $D_i$ denotes some $i$-th document.

Multiplicities are denoted by square brackets $[ ]$. Thus $[D]$, $[Q]$, etc., denotes, respectively, a multiplicity of documents, queries, etc. The number of component elements in a multiplicity is denoted by an exponent. Thus, $[Q]^k$ denotes a multiplicity of $k$ queries, where $k$ is some positive integer. When letters and numbers are written as superscripts to denote exponents in the usual mathematical sense, parentheses will be used to enclose such exponents. For example, $2$ raised to the $n$-th power will be written as $2^n$.

In accord with the usual convention, braces ${ }$ are used to denote sets. Thus sets are a kind of multiplicity for which rules as to membership have been or can be stated. Sets may be designated in a general fashion, i.e., $\{D\}$ indicates any set of documents while $\{D\}_j^i$ denotes the $j$-th set of documents $i$ in number.

Membership in a set may be specified by $\in$ in the usual way. Thus, $D_i \in \{D\}_k$ asserts that the $i$-th document is a member of the $k$-th set of documents. Inclusion may be denoted by $\subset$. Thus, $\{D\}_i \subset \{D\}_j$ means that all the members of $\{D\}_i$ are also members of $\{D\}_j$. The symbols $\in$ and $\subset$ may be used in the same way in speaking of multiplicities.

To indicate that a certain number of individual members constitute a certain multiplicity or set, the symbol $\ast$ may be used. Thus, if documents $D_i$, $D_j$, $D_t$, and $D_x$ constitute the $k$-th multiplicity of documents, we may indicate this by: $(D_i, D_j, D_t, D_x) \ast = \{D\}_k$.

Information retrieval methods are concerned so frequently with
some \( \{D\} \), with multiplicities of \( \{D\} \) denoted by \( \{\{D\}\} \) and with sets of \( \{D\} \) as denoted by \( \{\{D\}\} \) that the letter \( T \) is reserved to refer to a \( \{D\} \). Thus, instead of writing \( \{\{D\}\} \) we may write \( [T] \). Subscripts and superscripts are used with \( T \) to indicate, respectively, some individual class of documents and the number of members in the class; i.e., \( T^i_k \) denotes that the \( k \)-th class of documents consists of \( i \) documents in number.

It is sometimes convenient to use reverse braces \( } \) to indicate that we are thinking of a class as decomposed into its members. With subscripts we may then indicate some individual member; i.e., \( \{T^i_j\} \) would indicate the \( j \)-th document in the \( i \)-th class of documents. Similar use of \( ] \) is also possible; i.e., \( [D]^i_j \) denotes the \( j \)-th document within the \( i \)-th multiplicity of documents.

3.4 Notation for Relationships Involved in the Use of Symbols, Languages and Formats

The slant stroke will be used to indicate that a symbol pertains to some entity or concept, such as a document, class of documents, query, etc. Thus, \( S/D \) denotes a symbol assigned to some document. If we wish to indicate the \( k \)-th symbol assigned to the \( i \)-th document, we write \( S^k_i/D_i \). It may be the case that more than one symbol is associated with a single document. For example, \( [S^1_1, S^2_2, S^3_3, S^4_4]/D_1 \) indicates that the multiplicity of symbols \( S^1_1, S^2_2, S^3_3, S^4_4 \) is associated with the document \( D_1 \). We may also indicate such a relationship between some \( S \) and \( D \) by writing \( [S]/D \). Alternately, if we wish to infer that the symbols, associated with a document, may be selected by rule, we may
write \( \{S\}/D \). It is by no means uncommon for there to be a multiplicity of the kind of association notationally indicated by \([S]/D\) and \(\{S\}/D\), and this situation may be expressed by \([[S]]/D\) and \([[S]]/D\). We thus indicate that there is a multiplicity of associations in each of which a document on the one hand is associated with some \([S]\) or some \(\{S\}\) on the other hand.

Similarly, some one symbol may be associated with more than one document. For example, \(S_i/\[D_1,D_2,D_3,D_4\]\) indicates that the i-th symbol is associated with \(D_1,D_2,D_3,D_4\) which may be regarded as members of a multiplicity. In general, we may indicate that some symbol is associated with each document in a multiplicity by writing \(S/[D]\) and for some \(\{D\}\) or \(T\) we may write \(S/\{D\}\) or \(S/T\). The situation often occurs that there is a multiplicity of associations in each of which some symbol, on one hand, is associated with each member document of some \([D]\) or \(T\). Such situations may be specified notationally by writing \([S/[D]]\) or \([S/T]\) for sets of documents.

In presenting this notation for dealing with multiplicities and sets, it is not intended to rule out those multiplicities and sets that have only one member or none at all, often referred to as the null set.

Specifying how certain symbols are associated with individual documents, multiplicities and sets of the same may be accomplished with the aid of subscripts as illustrated by \(S_t/[D]_i\), which indicates that the t-th symbol is assigned to the i-th multiplicity of documents. Subscripts can also be used in analyzing situations in which a symbol, such as the word "cell" which has a number of meanings, is decomposed.
into individual codes for the various meanings of the word. If the symbol for "cell" is \( S_1 \), then various codes for cell would be designated by \( S_{11}, S_{12}, S_{13}, \ldots, S_{1f} \) if there are \( f \) such codes in all.

Correspondingly, we can indicate subdivisions of a \( [D]_1 \) by \( [D]_{11}, [D]_{12}, [D]_{13}, \ldots, [D]_{1f} \) where \( [D]_1 = [D]_{11} \cup [D]_{12} \cup [D]_{13} \cup \cdots \cup [D]_{1f} \), but it would not necessarily be the case that the submultiplicities would not have one or more member documents in common with one or more of its companion submultiplicities.

It is sometimes helpful to distinguish between two important ways in which a symbol or a part of a composite becomes associated with a document, query or the like. In one case, a symbol may be arbitrarily assigned e.g. to a document. Such assignment may be designated and consequently emphasized by an arrow pointing from the symbol to the entity, e.g., a document, to which it is arbitrarily assigned, e.g., \( S \rightarrow D \). The other case in which the symbol is derived by some process from an entity may be designated by an arrow from right to left, e.g., \( S \leftarrow D \).

It is sometimes advantageous to indicate that a given symbol, multiplicity of symbols, or set of symbols is expressed in a given language. Such relationship will be designated by the vertical stroke as illustrated by the general symbol \( (S|L) \). To designate the \( j \)-th symbol written in the \( k \)-th language, we write \( (S_j|L_k) \) and to denote some multiplicity of such symbols we write \( [S]|L_k \). The vertical stroke may be used in various combinations with the slanted stroke, as illustrated by \( (S|L_e)/[D]_1 \) to indicate a symbol in the \( e \)-th language for
the i-th set of documents.

It is sometimes the case that it is important to indicate the format of a language, symbol or the like. The general expression used is $(L|F)$. The symbolism $(L_j|F_r)$ would designate the j-th language written in the r-th format and $(S|L_j|F_r)$ would indicate a symbol so written.

To denote certain important categories of symbols and the like, we use lower case letters prefixed before the general symbol. Thus, any one symbol of the general category of symbols that refers to subject contents, especially of a document or query, will be indicated by $scS$. Classification symbols will be denoted by $clS$ and a subject content classification symbol will be designated by $scclS$.

3.5 **Notations for Component Parts of Composite Entities**

An individual dog may be regarded as a member of a class of animals, which class is often referred to by the term "dogs". The tail of an individual dog is certainly not a member of this class of animals. This example may serve to emphasize the importance of distinguishing between the whole-part relationship and the class inclusion relationship. Distinguishing between whole-part and class inclusion relationships is not always easy and obvious in documentation. Yet such distinction appears highly advisable as a safeguard against sloppy and falacious thinking. In some cases, such distinction is highly effective in directing attention to important distinctions.

The double vertical stroke $||$ is used to indicate that some, though not all, of the component parts of a composite entity are being
cited. For example, \( \{E\} || L_i \) denotes a set of elements, less than the complete set of elements, in the i-th language.

To denote that a set includes all of the parts of a certain kind, a triple vertical stroke is used. Thus, all the E's in the i-th language is designated by \( \{E\} ||| L_i \).

3.6 Notation for Processes and Operations

Perhaps the most important features of a process or an operation are the input and the output. The relationship between these factors will be represented in accordance with the following general scheme for a single process or operation: \( P \) (input \( \rightarrow \) output). We will, of course, encounter situations in which a process or operation acts on two or more inputs to provide two or more outputs. Such situations may be represented by appropriate modification of the general scheme. Namely, \( P \) ((input\(_1^1 + \text{input}_2 + \cdots + \text{input}_n\)) \( \rightarrow \) (output\(_1 + \text{output}_2 + \cdots + \text{output}_m\)), where n and m are not necessarily equal.

As discussed, an operation is regarded as a component part of a process. Hence, using the \( || \) to indicate the relationship of a component part, or an incomplete multiplicity, or an incomplete set of component parts to a whole, we may write such designation as follows: \( U_i || P_j \) to denote the i-th operation in the j-th process.

3.7 Conclusion

The notation presented in this paper has been quite deliberately developed as a shorthand means for describing what is observed when we direct close scrutiny to information retrieval methods and
documentation systems. Such description is considered an essential prerequisite to the investigation of this thesis. The notation has been developed in such a way as to minimize the possibility of prejudging the situation by applying one form or another of mathematical analysis and, in particular, to avoid overlooking important features of the information retrieval methods under consideration.
CHAPTER 4
ANALYSIS OF RETRIEVAL SYSTEMS

4.1 General

The starting point for analyzing information retrieval systems is postulating the existence of a file of documents, denoted by \([D]_m\), and of some query \(Q_1\) to which response is to be made by providing some subset of documents \(\{D\}_1\) where \(\{D\}_1 \subseteq [D]_m\).

A query is not regarded as existing until it is stated in some language and format. Accordingly, each query is expressed by some more or less complex symbol. Notationally, we may express such symbolic representation of a query as follows: \((S|L|F) \leftrightarrow Q_1\), where the language and format remain unspecified. With a given operational system, the query will be expressed in some specific language and format as notationally expressed by \((S|L_x|F_y) \leftrightarrow Q_1\). In servicing a query, it may have to be interpreted and translated. Consequently, specification of both language and format must be accorded due attention. It is, however, sometimes the case that we may wish to refer to a query symbol without specification of language and format as notationally indicated by \(S/Q\).

Likewise, in postulating the existence of a file of documents, each document can be said to exist only if its subject contents are symbolically expressed. Thus we arrive at the notational statement \((ttS|L|F) \leftrightarrow D\) where \(ttS\) designates the total text symbol for some
unspecified document D. In general information retrieval systems associate with a given document or with a set or multiplicity of documents one or more symbols for such purposes as serial numbering, class designation, characterizing for machine searching and the like. Such symbols may be notationally designated, in general, as follows:

\[ S|L|F \]^n \leftrightarrow D \]

for an individual document. To designate, in a general way, some symbol associated with some unspecified document, we will write \( S/D \). It is often necessary to refer to symbols of more or less complex nature that refer to the subject contents of some document or to some set or multiplicity of documents. Any symbol of this type will be denoted by \( scS \). A serial numeral or similar arbitrary identification symbol will be indicated by \( inS \).

In general, it will not be the case that the language (L) and format (F) are the same for the documents' symbols and for the individual queries to be searched. In order to perform searching the document symbols and query symbols must be interacted and, to this end, the expression of these two kinds of symbols may be required in the same L and F. If this be done, both the \( S/Q \) for each query to be serviced and the \( S/D \) for the documents in the file to be searched are constituted from the same domain of elements, or in notation this means that \( \{E\} || S/Q = \{E\} || L_z | F_w \) and \( \{E\} || S/D = \{E\} || L_z | F_w \), where \( L_z \) denotes the common language and \( F_w \) some common format for \( L_z \).

Any search involves bringing about some kind of selective interaction between \( S/Q \) and the totality of individual \( S/D \) for a given file. When \( L_z \) is a human language, e.g., English, and this selective interaction is performed by people, a complete logical analysis
of the interaction is probably impossible. Such analysis might well be even more difficult than the analysis of the complex processes performed by a human in translating from one natural language to another. However, \( L_z \) may be some specially defined language which may or may not have been derived from a human language. In such a case, detailed analysis and specification of the interaction may be possible.

Various information retrieval systems conduct preparatory processing of documents so as to generate symbols that constitute sequences of elements that are recorded in a form that can be detected by more or less complex searching and selecting machines. In such systems as discussed here, each document may be regarded as having an associated scS/D which is built up of elements selected as appropriate from the totality of elements that are provided by the information retrieval system for characterizing the subject contents of documents. Thus, for each document, we will have some \( (\{E\}|||scS) \leftrightarrow \text{D} \), and prior to conducting a search the query will be expressed in terms of a set of elements \( (\{E\}|||S) \leftrightarrow Q \). When both \( (\{E\}|||scS) \leftrightarrow \text{D} \) and \( (\{E\}|||S) \leftrightarrow Q \) are expressed in some well-defined common language \( L_z \) and common format \( F \), an automatic device may be used to examine the relationships between such symbols. The definition of a set of elements that specifies a given \( Q \) is not restricted to the logical intersection of elements as illustrated by \( E_1 \cap E_2 \cap \cdots \cap E_k \). The logical union may also be involved as illustrated by \( (E_1 \cup E_2 \cup E_3) \cap (E_3 \cup E_4) \) as well as the logical difference, e.g., \( E_2 \setminus \overline{E_4} \). Furthermore, a multiplicity of levels of such logical specifications may be established. In the ASM-WRU system, logically defined combinations of semantic factors may define
classes of terms, code phrases may be defined as logical combinations built up of classes of terms either with or without role indicators, code sentences may be defined in terms of logical combinations of phrases, and code paragraphs and entire encoded abstracts as logically defined combinations built up of code sentences and code paragraphs. But, no matter how simple or complex such specification may be, the searching as performed by the automatic equipment determines which of the following two situations prevails, namely: \((\{E\}|||S)/Q \subset \mathcal{C} \lor (\{E\}|||S)/D\). When the inclusion relationship exists, we may properly conclude that the document corresponding to the S/D is of pertinent interest. When the non-inclusion relationship exists, it is the case that the symbol for the document does not comprise all the elements in S/Q as set up to indicate the scope of the query in question.

This later condition may come about in more than one way however. First and most obvious, the document in question, even if its total text form were to be translated into our language L, would not include all the elements as specified in \((\{E\}|||S)/Q\). In such a case, the document in question may or may not be of interest to the query, for it is easy to imagine a situation in which \(D_1 \cup D_2 \cup \cdots \cup D_k\) would generate a resultant or composite document \(D_r\), whose subject contents would provide a \((\{E\}|||S)/D_r\) which would include within its scope the \((\{E\}|||S)/Q\). This situation might well exist even though the inclusion relationship did not hold for any of the individual documents \(D_1, D_2, \ldots, D_k\).

Even if the document \(D_r\), so generated, did not exhibit the inclusion relationship with respect to the given query, the various documents
comprising $D_r$ may perhaps still be of pertinent interest. For example, the addition of some one document $D_a$ unioned with $D_r$ might generate a $(\{E\}||S)/D_{rwa}$ which would exhibit the inclusion relationship. In other words, the analysis of the interrelated query elements and of the documents available might well point the direction of further research and experimentation. This principle has been applied in intelligence work, when the Essential Elements of Information are known. Existing information is compared to the EEI and then efforts are made to obtain those elements that are missing, in order to complete the intelligence picture. This same principle often guides industrial research and development activities. It should be noted, in this connection, that such piecing together of scattered information is far less important in novelty searching as conducted by the Patent Office. An effective reference against an application must be close enough to the latter's subject content that bridging the gap would require no more than the average competence of a person skilled in the art.

The non-inclusion condition may also occur even if the ttS/D does include all the elements specified in $(\{E\}||S)/Q$ provided that in converting the ttS/D into our scS/D some of the elements as found in our S/Q were omitted. In this case we would speak of a critical loss of information during the input processing. It is obvious therefore that the fewer the elements contained in the S/D, the more difficult it will be to select the elements so as to avoid such critical loss of information. It is extremely important in designing an information retrieval system for Patent Office novelty searches to avoid such
critical loss of information when processing documents to generate a subject content symbol. When dealing with scientific and technical publications, it is usually the case that the $ttS/D$ for each document provides a sufficient range of information so that it would be possible to generate, for a given $D$, a $(\{E\}||scS)\nsubseteq D$ which would be inclusive with respect to the $(\{E\}||S)\nsubseteq Q$ for realistic queries, e.g., those generated by Patent Office examiners in conjunction with novelty searches. It is equally essential that the $(\{E\}||scS)\nsubseteq D$ shall be set up in such a way that the inclusion relationship does not occur for many of those documents that are not of pertinent interest to the queries being searched.

Thus, it may be concluded that the central problem in designing an information retrieval system to meet the novelty searching requirements of the Patent Office is to establish methods to generate, for each $D$, an appropriate $(\{E\}||scS)\nsubseteq D$. Resolving this problem will require:

1. Establishing a code system, i.e., a language $L$, whose elements enable both the subject contents of documents and of queries to be expressed so that interaction between the various $(\{E\}||scS)\nsubseteq D$ and a given $(\{E\}||S)\nsubseteq Q$ will provide the desired selective discrimination.

2. Formulating procedures for generating an appropriate $scS/D$ for each $D$ and for expressing each query as an appropriate $S/Q$.

3. Ensuring that costs do not rise to excessive levels by working out methods that attain efficient use of human effort and also of automation techniques to perform large volumes of routine work.
The only classification system considered here will be that used in the United States Patent Office. In this system, the documents are physically divided into subclasses \( T \) and there exists a multiplicity \( K \) of these \( T \)'s. It is the case in this system that we are dealing with hierarchical sets which are characterized by inclusion relationships. In considering any one of the 311 major classes of the patent system, we will regard it as being of the first echelon and for such an inclusion class we will use the symbolism \( T \) and for the major subclasses, the first indentation in the Patent Office classification schedule, we will use the symbol \( T \) to denote any one of such second echelon subclasses. In general, the hierarchical sets in the Patent Office classification system are so organized that each subclass of the \( n+1 \) echelon must be regarded as a subset within some one more inclusive class of the \( n \)-th echelon, or more concisely, \( n+1 \subseteq n \). The highest value of \( n \) in \( T \) is not uniform throughout the Patent Office classification system but values of \( n \) larger than six are rarely observed. A class number is assigned so as to designate uniquely each class and subclass, and also to specify the hierarchical relationships of classes and subclasses as organized at various echelon levels. Such a numerical symbol may be designated notationally by incl\( S/T \). Each such numerical class designator is accompanied by a subject heading whose interpretation requires that we remember that each such heading, e.g., at echelon \( n+1 \), distinguishes a given subclass from its immediately inclusive subclass, e.g., at echelon \( n \).
Thus, the subject headings for a given subclass must be taken as in-
cluding the subject headings of all the successive inclusive T's whose
echelon index numbers diminish as we take account of the successive
inclusive T's. Consequently, the subject content symbols for a given
T, which notationally we may designate by scclS/T, must be regarded
as a composite which includes the subject headings for its successive
inclusive T's as presented by the classification schedules.

The intricacies of arranging T's in echelon, the assignment of
numbers to the T's to indicate their hierarchical ordering and the need
for care in determining what the scclS may be for a given T should not
cause us to overlook the fact that conventional classification designates by various symbols, our scclS/T, certain subject content characteristics of the documents within the given T. Furthermore, each such
scclS/T consists of a combination of language elements. In other words,
the classification provides a multiplicity of class symbols, any one of
which may be denoted by (\{E\}|||scclS)/T. These symbols are decisively
important in using the Patent Office classification system to meet
novelty search requirements.

In analyzing further the role of classification symbols let
us note, first of all, that each such classification symbol may be re-
garded as referring to every member document in the given class. As
a consequence the use of conventional classification to service a given
query involves interaction between (\{E\}||| scS)/Q and the (\{E\}||| scclS)/T
as provided by the classification system. Such interaction is facili-
tated by such aids as the following: (1) The hierarchical arrangement,
which often permits a given major class to be passed over as uninteresting by doing no more than considering the general class heading. (2) A subject index, which is an alphabetized array of subject content symbols $[\text{scixS}] ||| \text{ixa}$. Each such scixS refers to one or more T's as may be indicated by $\text{scixS}/[T]^n$ where $n \geq 1$. (3) Classification bulletins which define the meanings of various terms used in the classification headings, i.e., of various E's in individual $(E \text{ scclS})/T$. These bulletins also present and explain policies and rules for assigning patents to various T's.

As a consequence of such interaction between a given $(\{E\} ||| \text{scS})/Q$ and the various $(\{E\} ||| \text{scclS})/T$, aided perhaps by use of the subject index, one or more T's are selected as containing patents of possible pertinent interest for the Q in question. The Patent Office classification system functions, in other words, by directing attention to pre-established T's rather than to specific documents or to some $\{D\}$ whose members are determined by interacting a $(\{E\} ||| \text{scS})/Q$ with subject content symbols for individual documents as denoted by $(\{E\} ||| \text{scclS})/D$. Personnel inspection is necessary to determine which documents may be of pertinent interest in the $[T]^n$ to which the classification system directed attention. The novelty searching process, when conducted by means of the conventional classification system may be regarded as consisting of two operations: (1) Determining which $[T]^n$ may contain documents of pertinent interest to deciding the patentable novelty of an application. (2) Personally examining the documents in the $[T]^n$ to determine which are indeed of pertinent interest. (We will refer to a document of pertinent interest by the symbol Dp
Symbolically, these two steps may be represented as follows:

(1) \( \text{csU}(\{\{E\}\ || \ scS)/Q + (\{E\}\ || \ scclS)/T) \rightarrow \{cLs/T\}^N \) where \( \text{csU} \) indicates the operation of searching the headings of the classification system. (2) \( \text{psU}(\{T\}^n \rightarrow \{Dp\} + \{Du\}) \) where \( \text{psU} \) indicates the operation of personally searching the documents that are members of \( \{T\}^n \).

As the symbolic notation makes clear the output of the first step is a series of symbols that identify various \( T \)'s and such symbols enable appropriate \( T \)'s to be selected as input for the second step.

It is highly desirable, as our analysis indicates, and as rising costs of novelty searches in the Patent Office fully confirm, that the costs for performing the second of these two steps to be kept down. As a matter of fact, these costs have become intolerably high and it seems clear that this upward trend can be reversed only by taking drastic action. A somewhat more detailed analysis of the two general steps above stated is, accordingly, in order.

It may be noted, first of all, that the output of the second step may be regarded as being made up of two sets of documents, which may be further characterized by their \( x \) and \( y \) superscripts in \( \{Dp\}^x \) and \( \{Du\}^y \). When \( x < y \), much personal effort and time must be devoted to eliminating from the members of \( \{T\}^n \) those documents of no interest. In striving to minimize the value of \( y \), two requirements must be satisfied: (1) The individual document classes to which step one directs attention must not consist of a large number of documents. In other words, for each \( T \) in the \( \{T\}^n \) the value of \( k \) in \( T^k \) must be kept at a relatively low level. In practice, this means that those \( T \)'s which
are inclusive of higher echelon T's are to be avoided as members of \( \{T\}^n \). Stated positively, the members of \( \{T\}^n \) should be those T's that comprise the smallest number of documents. (2) The value of n in \( \{T\}^n \) as selected by the first step must be kept down as otherwise the advantages of keeping low the number of documents in the T's, especially the higher echelon T's, are lost.

The first requirement can be met, with increasing number of documents embraced by a classification system, only by increasing the number of subclasses. In other words, as the total number of documents in a conventional classification system increases, a larger number of T's is required. Either higher order echelons must be set up or, if the number of echelons is held constant, a larger number of T's at the various echelon levels is required. The number of characteristics, our E's, that are involved in distinguishing between the various T's increases inevitably in either case.

With increase in the number of E's, the number of different possible combinations among them also increases but much more rapidly. Each of such possible combinations is potentially both a subject content classification symbol and also a subject content query symbol; but conventional classification permits only a fraction of the totality of potential subject content classification symbols to be actually used as sccls/T as the costs of setting up and maintaining a very large number of different T's otherwise becomes excessive. A further adverse factor has already begun to become apparent even when the number of T's is held to a level far below the number of the subject content classification symbols that might be generated from the set of E's in
actual use. This factor is the effort and cost of conducting our csU. If the total number of such subject content classification symbols were to be greatly increased, then the cost of performing csU can be expected to rise to the point that the use of automatic techniques would become promising to say the least.

Thus the use of pigeon holes or other fixed compartmentalizing systems have the following inherent disadvantages when dealing with large numbers of documents whose subject matter is complex: (1) The number of T's and, hence, the corresponding number of subject content symbols is restricted by the practical costs of maintaining large numbers of T's. (2) The probability that all the members of $\{Dp\}$ for a given query will be found within a single T decreases as the system becomes more ramified and, hence, $y$ in $\{Du\}_y^x$ steadily increases with respect to $x$ in $\{Dp\}_y^x$.

It should be emphasized that these conclusions are not theoretical derivations but statements of what has occurred as the Patent Office system has evolved during the past century. Confirmation is provided by rising searching costs and the benefits attained by mechanized searching as reviewed in section 2.5 of this thesis.

Analysis of the manner in which conventional classification has become increasingly less satisfactory for Patent Office novelty searching indicates that there are two general steps required in order to develop effective means for coping with the problem: (1) Developing a list of appropriately selected $E$'s from which to develop $\{E\}_k || scclS)/Tk$ where $k \geq 1$. (2) Provision of means for selectively interacting the $\{E\}_k || scclS)/Tk$ with a $\{E\}_k || scS)/Q$ so as to provide
a \( \{T\}^n \) which comprises most if not all the pertinent documents together with a tolerably small number of unwanted documents, where as before \( k \geq 1 \).

Here \( k \) may be relatively large for the more inclusive \( T \)'s, provided that personal review of large numbers of documents can be avoided. This becomes possible if machine searching operations can be directed to the subject content symbols that pertain to small numbers of patents. In the extreme case, \( k \) may be equal to 1 at the highest echelon. Such a case would correspond to grouping patents into a smaller number of \( T \)'s than the present approximately 57,000 subclasses but using a limited degree of echelon classification to select sets of recordings of \( (\{E\} \cap \{scclS\})/D \) for machine searching.

It is from this point of view that previously developed information retrieval systems will be reviewed and evaluated.

4.3 Uniterm System

In this system the analysis of documents results in generating for each document a set of uniterms, each of which is a subject content symbol. This we may designate by \( \{scuS\} \leftrightarrow D \). An identifying symbol is also associated with each document as denoted by \( inS 
\rightarrow D \).

In general, it is the case that any one uniterm is associated with more than one document as denoted by \( scuS \leftrightarrow \{D\} \).

A further essential feature of the uniterm system is that each uniterm is recorded in conjunction with the corresponding set of identifying symbols for the member \( D \)'s in its associated \( \{D\} \). Thus we arrive at a composite uniterm file entry, which we may denote by \( feuS \).
and which consists of \( scuS \Leftrightarrow \{D\}_1 + \{inS/D\}_1 \). Such a uniterm file entry must be generated for each uniterm in a given system.

Searches performed when using the Uniterm system consist of the following operations: (1) Deciding what combination of uniterms correspond to a given query, i.e., generating a \( \{scus\} \not\subseteq Q \). (2) Interacting that set of \( feuS \) whose \( \{scuS\} \equiv \{scuS\} \not\subseteq Q \) so as to determine a \( \{inS/D\} \) each of whose member \( inS/D \) occurs in each member of the \( \{feuS\} \). As stated here, only those individual documents would be selected whose \( \{scuS\}/D \supset \{scuS\} /Q \) when the inclusion relationship is defined solely by the logical product. In theory, the logical sum and difference could be used in specifying this inclusion relationship. In practice, this extension of the definition of the inclusion relationship has been done rarely, if at all, as the required interaction among members of \( \{feuS\} \) becomes tedious to perform clerically by personnel and it requires relatively extensive and time consuming computer programs. (3) Personal inspection of the \( \{D\} \) whose members have been identified by their corresponding identifying numbers that are members of the \( \{inS/D\} \) above referred to. This operation can be notationally designated by \( psu((\{D\}_n \rightarrow \{Dp\}_x \uplus \{Du\}_y) \) with the desirability as before that \( y \) in \( \{Du\}_y \) shall be held to tolerably low values.

The limitations of the uniterm system which are either obvious or implicit may be summarized as follows: (1) The system permits, in practice, only limited exploitation of the logical sum and logical difference. (2) In addition to this restriction, only one level of Boolean logic is available to define queries, that is, to specify a \( \{scuS\} \) for a given \( Q \). (3) The recording of syntactical
language elements, though possible in theory, encounters such practical difficulties that such elements can be used, if at all, only to a very limited degree. This becomes a particularly important drawback when dealing with large files of complex subject matter, as syntactical relationships can provide much useful discrimination. (4) Previous more general review of the uniterm system has pointed out that entering a new in$S/D$ under each feu$S$ for the members of $\{scuS\}/D_i$ can become a cost factor of by no means negligible proportions if the number of uniterms in the system is large and if the value of $n$ in $\{scuS\}^n/D$ is not held to very low values. (5) Well-defined uniterms have been found to be a virtual necessity if a uniterm system is to function effectively. This requirement is relatively easy to meet when working in a narrow field but with broad ranges of subject matter, severe difficulties are encountered unless the list of uniterms is sharply restricted. (6) If close control on the definition and use of uniterms is not maintained, it becomes difficult to establish a $\{scuS\}/Q$ that is effective in directing attention to a full set of pertinent documents in a given file.

4.4 Descriptor System

As indicated by previous discussion, this system has much in common with the uniterm system. Input analysis of a document results in generating a $\{scdS\} \not\in D$ where scd$S$ indicates a descriptor which may also be regarded as an $E$ in this system. Furthermore, an identifying number or the like is associated in this system with each document.

In contrast to the uniterm system, the entire set of descriptors
for each document is recorded as a symbol sequence or symbol grouping together with the inS for the document. This system was designed to work with very simple equipment and for many situations this is, of course, an important practical advantage. However, for situations in which the capabilities of more elaborate electronic equipment may be applied to advantage, the descriptor system's very simplicity may be a serious disadvantage as discussed subsequently in more detail.

Searches performed when using the descriptor system consists of the following operations: (1) Deciding what combination of descriptors corresponds to a given query, i.e., generating a \( \{\text{scdS}\} \not\in Q \).

(2) Interacting the \( \{\text{scdS}\} \not\in Q \) with the \( [\{\text{scdS}\} \not\in D] \) as provided by the system to select a \( \{D\}^n \). Descriptor systems have been worked out so that simple apparatus for sorting edge-notched punched cards can be operated so that the above indicated interaction can be performed simultaneously for a considerable number, say 100 or more cards, each of which records one \( \{\text{scdS}\} \not\in D \). However, such cards do impose severe restrictions on the number of scdS in any one \( \{\text{scdS}\} \not\in D \). Descriptor systems that make use of magnetic tape recording are not subject to this restriction. (3) Personal inspection of the \( \{D\}^n \) as indicated by the notational expression, \( \text{psU}(\{D\}^n \rightarrow \{Dp\}^X \ast \{Du\}^Y) \).

The principal limitation of this system is its use of a restricted set of descriptors, or E's in our notation, for the analysis of documents to generate a \( \{\text{scdS}\} \not\in D \) for a given document. The consequence is corresponding restriction in the \( \{\text{scdS}\} \not\in Q \), that is to say, it is all too often the case that a \( \{\text{scdS}\} \not\in Q \), if set up so that step
2 above will select all or nearly all the pertinent documents in a file, will also result in simultaneously selecting a \( \{D_u\}^y \) with an intolerably high value of \( y \), when dealing with large numbers of documents. This leads, of course, to corresponding unacceptably high costs for performing the personal inspection of the selected \( \{D\}^n \) as summarized in step 3 above.

It is also to be noted that this system, as originally developed, applied devices that provided the ability to define selecting operations in terms of a single level of Boolean algebra. This restriction may be surmounted when an electronic computer is used however.

4.5 **ASM-WRU System**

To provide a common basis for comparing this system with those previously discussed, it may first be noted that this system generates for each document a subject content symbol for machine searching, as notionally designated by \( \{scmsS\} \leftarrow D \). Each such symbol consists of a set of elements as we may indicate by writing \( \{E\} \parallel \{scmsS\} \leftarrow D \). The manner of selecting and organizing the \( \{E\} \) in constructing the \( \{scmsS\} \leftarrow D \) for a given document is a matter to be considered subsequently.

As recorded for machine searching, each \( \{scmsS\} \leftarrow D \) is associated with a corresponding \( inS \rightarrow D \) for the document in question. Up to the present, this system has been operated by recording file entries composed so that each such entry consists of a pair of \( \{E\} \parallel \{scmsS\} \leftarrow D \) and \( inS \rightarrow D \) for some one document.
Searches performed when using the ASM-WRU system consist of the following operations: (1) Encoding a query as a $(\{E\} || scmsS)\not\leftrightarrow Q$. The elements used and the rules for generating a $scmsS \not\leftrightarrow Q$ are the same as for the encoding of documents. (2) Interacting the $(\{E\} || scmsS)\not\leftrightarrow Q$ with the $[(\{E\} || scmsS) \not\leftrightarrow D]$ for the file of documents to be searched. This selective interaction can be performed at a low cost by appropriately designed and programmed automatic electronic equipment. In this way a set of documents $\{D\}_{n}$ is selected as being of probable pertinent interest. As previously noted in section 2.5, syntactical patterns may be taken into account when generating a $(\{E\} || scmsS)\not\leftrightarrow Q$ and selectively interacting such a symbol with the members of $[(\{E\} || scmsS) \not\leftrightarrow D]$. Multi-level Boolean algebra provides the general abstract framework for generating the various symbols involved and for specifying their interaction. (3) Personal inspection of the $\{D\}_{n}$ whose members have been identified as outlined above. This operation can be notationally designated by $psU(\{D\}_{n} \rightarrow \{dp\}_{x} + \{Du\}_{y})$. As with other encoding systems, it is highly desirable that $y$ shall be held at low values and that $\{dp\}_{x}$ shall contain the acceptably large percentage of pertinent documents.

As this outline and previous general description of this system may have made clear, it makes use of various substantive terms to characterize the subject content of documents. As with the uniterm and descriptor system, such substantive terms may designate substances, properties, processes, conditions, persons, organizations, locations and the like. The ASM-WRU system differs from the two previously
reviewed systems in the following respects: (1) Codes for individual substantive terms are organized into syntactical patterns with the aid of role indicators. The basic units in such patterns are combinations of at least one role indicator with the code for an individual substantive term to generate phrases or, more precisely, the analogs of phrases. Such phrases are recorded for machine searching as units whose beginning and end are indicated by symbols that the searching machine is able to detect and take into account when selecting (or rejecting) a document as pertinent (or unwanted) for a given search. Furthermore, phrases are combined into sentences, the latter into paragraphs and the paragraphs into messages. The rules for building up such syntactical structures may be varied to accommodate the system to a given set of requirements and circumstances. (2) The code for individual substantive terms are constructed in such a way that their relationship to certain members of a set of selected generic terms is indicated by code elements. More specifically, each code for an individual substantive term consists of a small number, usually five or less, of code elements each of which specifies a generic term, for example, "machine or device," "process of connecting," "a general class of substances such as alloys." Each of the code elements or semantic factors is accompanied, in codes for specific terms, by an auxiliary symbol that indicates how the specific term is related to the semantic factor as included in the code. Thus, the code for a device used for welding stainless steel will be constructed so as to indicate that the device is a member of the general class of "machine or devices," that
it performs the "process of connecting" and acts on "alloys." A numerical suffix is provided as the final part of each code for a specific term to insure that each code is unique for every term in the code dictionary with exception of complete synonyms, such as mercury and quicksilver. When encoding a query preparatory to searching, both individual terms as well as sets of terms whose codes contain an individual semantic factor or some combination of the same may be specified.

Summarizing, we may say that the designation of syntactical patterns and the generation of codes that relate specific terms to generic concepts are features that distinguish this system from the previously reviewed uniterm and descriptor systems. These features extend the set of elements that are available for generating both $(\{E\}|| scmsS)/D$ and $(\{E\}|| scmsS)/Q$ and thereby make possible more discriminating and flexible searching. These capabilities appear highly useful in developing specific methods for coping with the novelty search requirements of the Patent Office. Accordingly, this system was selected to conduct exploratory studies as presented in subsequent chapters of this thesis.
CHAPTER 5
ANALYSIS OF BENEFITS AND COSTS

5.1 Introduction

This investigation has been conducted in the conviction that the novelty search problem of the Patent Office cannot be realistically analyzed unless due attention is accorded to economic factors, in particular, to the benefits and costs of meeting the Patent Office search requirements by applying some method or combination of methods. Such an analysis of the Patent Office operations can be divided into two parts, out-office and in-office benefits and costs.

5.2 Out-office Benefits and Costs

As the introductory review of our patent system pointed out, it benefits the nation by encouraging invention, technological innovations and industrial progress. Prompt publication of new technological information by avoiding delay in issuance of patents is a most important factor in securing the benefits that the patent system has provided in the past. These benefits include not only encouragement to inventors, stimulation of industrial activity as fostered by grant of patent protection, but also informing the industrial community of new technological advances. The benefits of speeding up performance of novelty searches to avoid delay in issuance of patents are very real and important though their precise evaluation and expression in
terms of dollars are admittedly difficult.

The out-office costs are those connected with the delay in issuing and publishing patents. Under United States law, the contents of a patent application are considered as secret and cannot be disclosed to anyone. Consequently, until a patent is actually issued the public is denied access to the new knowledge contained therein, and as noted previously such delay often amounts to three or four years at present. Cost factors involved here include (1) the cost to industry of duplicating, or at least attempting to duplicate, the technological advance described in patent applications undergoing examination and (2) the cost to the nation's productivity of not having available the latest developments of science and industry. These costs, although hard to evaluate, are very real costs to the country and cannot be disregarded. Another closely related cost is the loss to the applicant of revenue from the use or sale of his invention prior to the issuance of the patent. Delay in issuing a patent may delay or even dissuade business management from making investments required to introduce an invention into industrial operations. In such a situation subsequent reinvention by another firm or individual may lead to confused and difficult situations in which the original inventor and his employers are unable to proceed with exploitation of an invention because of delay in issuance of a patent. If another firm or individual should make the same discovery after the applicant, but prior to the issuance of the patent, he may use it or sell it during this period with no return to the original inventor. Although such costs never
appear in the budget of the Patent Office, they are costs to the nation and, therefore, to the government of which the Patent Office is a part. Such costs are a function of delay in the Patent Office procedures and, consequently, are directly related to the Patent Office costs occasioned by slow procedures. An increase in the costs of such procedures could be justified if, by faster action in the Patent Office, the above outlined out-office costs could be drastically reduced. It would, of course, be even more desirable to develop methods that would reduce both out-office and in-office costs simultaneously.

5.3 In-office Costs and Benefits

5.3.1 General Statement

The in-office costs of the Patent Office are directly related to the servicing of queries as generated by patent applications. When an application is received it is reviewed initially as to form for compliance with the rules of the Patent Office and then forwarded to an examining division and after that, to a patent examiner. Regardless of the searching system in use, the patent examiner by preliminary analysis of the application generates the actual queries to be serviced by searching. The system in use for retrieval may influence the scope and form of the queries, but their generation by analysis of the application remains basically the same. This generation of queries may well be thought of as the prime initial function of the examiner. Once appropriate queries have been framed and correspondingly selected documents have been provided to the examiner, discharge of his main responsibility, namely, evaluating the application as to inventive novelty,
becomes possible.

The searching operations, subsequent to initial formulation of one or more queries, constitute the principal field of this investigation, namely, the application of information retrieval methods to search previously issued United States patents to meet novelty search requirements. The objective of the novelty search is to place on the desk of the patent examiner a set of patents which conforms to two basic requirements. In the first place, the set of patents should include all previously issued United States patents of pertinent interest, for if this is not achieved the examiner's evaluation of novelty may leave out of consideration some previously issued patent that should be taken into account. Secondly, the set of documents must not contain a large number of non-pertinent or marginally pertinent patents. The elimination of such patents by personal inspection has become excessively costly when conventional classification is used as the principal means for providing examiners with sets of patents for evaluating various applications.

Accordingly, the searching phase starts with the formulation of one or more queries by an examiner as a result of preliminary review of a patent application. Next, the query or queries is brought into interaction with the documentation system to provide some set of patents as may be designated notationally by \( \{D^N\} \). The search phase is concluded by the examiner reviewing the member of \( \{D^N\} \) to separate out the pertinent items \( \{D_p\}^X \) from those of no pertinent interest, \( \{D_u\}^Y \). In conducting a search, there are, accordingly, two major cost
areas, once the examiner has generated an appropriate query or queries by preliminary review of an application. These cost areas are (1) operating the documentation system to provide a $\{D\}_n$ for a given $Q$ and (2) inspection of the $\{D\}_n$ to isolate the $\{Dp\}_x$ from the $\{Du\}_y$.

In order for a documentation system to respond to a query $Q$ with a set of documents $\{D\}_n$, the system must, of course, be set up and in existence. In this connection there are three major areas of costs: (1) Developing the documentation system and providing needed equipment. (2) Processing documents into the system, in particular, classifying, indexing, encoding or otherwise characterizing the documents subsequently to be searched. (3) Training personnel to perform various tasks to operate the documentation system.

It is particularly important to note that a major investment is required in order to establish a new documentation system on an operational basis. Although costs for equipment and for training personnel are not to be overlooked or underestimated, the major investment is almost certain to be in the analysis and related input processing of extensive files of documents, i.e., of previously issued United States patents for the situation under consideration in this thesis. It should also be noted that, as new documents are processed into the system, further input processing costs are incurred.

In working out the design of a system to meet the novelty search requirements of the United States Patent Office, it is instructive to consider how the capabilities and limitations of various previously reviewed documentation methods have been found, or may be
expected, to influence the above-outlined overall costs of meeting
novelty search requirements.

In evaluating overall costs, it is necessary to take account,
as outlined above, of three general types of costs: (1) Cost of de-
signing a system and providing equipment and personnel. (2) Costs of
operationally establishing and maintaining a given documentation sys-
tem. (3) Costs of operating the documentation system to meet the re-
quirements of the situation in question.

Minimizing the sum of these three types of costs, while meet-
ing the requirements of the situation, is the objective of designing
information retrieval systems.

5.3.2 Conventional Classification System

(1) Cost of designing the system and providing equipment and
personnel. Previous Patent Office application of conventional clas-
sification is so extensive that system design costs and personnel
training costs may be regarded as already amortized. The costs of
the equipment used, namely, pigeon holes or similar arrangements of
compartment, are very low.

From time to time, parts of the classification system require
revision in order to provide a new set of subclasses of those classes
into which excessively large number of patents have been classified.
There is, accordingly, a continuing system design cost that must be
taken into account.

(2) Cost of operationally establishing and maintaining the
system. The present conventional classification system embraces all
previously issued United States patents and this system constitutes, as it were, a major investment on the part of the Patent Office. In designing a new documentation system, it would, obviously, be most advisable to take full advantage of this previous investment.

Maintaining the present conventional classification system involves two different, though related, activities: (a) Newly issued patents are continually incorporated into the previously established system of classes and subclasses. (b) When, as a result of the continuing incorporation of newly issued patents into the previously established system, various subclasses include an excessive number of patents, the system must not only be revised, as outlined above, but the patents in such subclasses must be assigned to newly defined subclasses. The costs thereby incurred are both considerable and recurrent.

(3) Cost of operating the documentation system. Once a given query has been stated, costs are incurred by the following: (a) A decision must be made as to which subclasses may contain patents of pertinent interest to the query. Though not excessive at the present time, costs of this kind tend to increase as the conventional classification system is made more ramified. (b) The subclasses identified as possibly containing patents of pertinent interest to the examiner must be supplied or made available to him. The costs involved here pertain to such factors as the following: (i) providing files and office space so that examiners may have frequently consulted subclasses directly accessible in their offices, (ii) maintaining
classified sets of patents to be provided to examiners as needed to supplement such files as may be maintained in their respective offices.

(c) Personal examination by the examiner of the patents in the subclasses to which the classification system has directed his attention. Excessively high costs in this connection have made it imperative for the Patent Office to develop a documentation system whose discriminating capabilities are much superior than those attainable by conventional classification.

5.3.3 Uniterm System

(1) Cost of designing the system and providing equipment and personnel. The development of the best possible set of uniterms could be approached by first taking account of the terms used in the subject headings of the previous conventional classification system. It seems likely that a listing of such terms would have to be extended very considerably, in particular, to include more specific terminology, if the unterm system were to avoid being inadequate because of insufficient reliability in selecting pertinent patents and rejecting unwanted documents during the searching phase.

It would also be advisable, and virtually necessary, to organize the listing of uniterms in such a way that they might be used in a consistent fashion both during input processing of documents and also when stating what combinations of uniterms should be used to effect a search in response to a given query. The amount of time and effort required for generating such listings of uniterms could be expected to be extensive, perhaps surprisingly so, as much more is
involved than merely picking up individual terms as they are encountered in previously issued patents.

Personnel would have to be trained to perform the following kinds of tasks: (i) input analysis of individual patents and characterization by an appropriate set of uniterms, (ii) posting of patents under the various uniterms, (iii) interpreting examiner's queries as corresponding sets of uniterms, (iv) conducting searching operations by operating the unterm searching system to select a set of patents to respond to a given query.

Equipment requirements would pertain to the posting and searching operations as outlined under (ii) and (iv) above. Providing listings of uniterms and the updating of such listings should also receive attention in connection with equipment requirements.

(2) Cost of operationally establishing and maintaining the system. At present each patent is assigned to at least one subclass within the conventional classification system of the Patent Office. It seems very doubtful that an effective unterm system could be established by doing no more than assigning to each patent a set of uniterms obtained from the subject headings of the subclasses to which it is now assigned. Rather, it must be anticipated that the subject contents of previously issued patents would have to be reviewed in order to provide more detailed characterization to provide sufficient discrimination to avoid excessive costs as outlined in the next section.

In setting up a unterm system, the posting of document identifying numerals under the various uniterms is a cost area that must
be taken into account. For large numbers of documents, automation techniques might be applied to advantage to reduce the costs of such posting.

As new technological concepts develop, it seems likely that extension and revision of the list of uniterms would be a virtual requirement. Furthermore, it might also prove necessary to revise the sets of uniterms previously assigned to various patents. The extent to which this would prove necessary would depend on the course of technological innovation, and it is not likely to be easy to predict future costs in this regard.

(3) Cost of operating the documentation system. Previous analysis of the uniterm system as presented in section 4.3 identified the following operations, with each of which operating costs would be associated: (a) Deciding what combination of uniterms corresponds to a given query, i.e., generating a \( \{ \text{scuS} \} \not\leftrightarrow Q \). (b) Interacting the \( \{ \text{scuS} \} \not\leftrightarrow Q \) with the recordings of the uniterm system to generate a listing of identification numbers for patents selected as being of possible pertinent interest. (c) Providing to the examiner copies of the patents identified as being of possible pertinent interest, i.e., providing him with a \( \{ \text{D} \}^n \). (d) Personal examination by the examiner of the \( \{ \text{D} \}^n \) to select those patents of pertinent interest to the application undergoing examination. The limitations of the uniterm system as previously discussed in section 4.3 make it very doubtful that this system can provide sufficient discriminating power so that excessive costs could be avoided in this connection.
5.3.4 Descriptor System

(1) Cost of designing the system and providing equipment and personnel. The development of the best possible set of descriptors would be closely similar to working out the best possible set of uniterms. This is a direct consequence of previously reviewed similarities between the uniterm and descriptor systems. To provide a descriptor system capable of providing the required measure of discriminating power, the degree of characterization of the subject contents of each patent seems certain to need to be more detailed than has been the case in many descriptor systems as developed to deal with narrow ranges of subject matter. As a consequence, to provide greater discriminating power for Patent Office purposes, higher costs than in dealing with narrow ranges of subject matter would have to be anticipated in systems development as follows: (i) establishing the descriptors, (ii) providing equipment for implementing the descriptor system. (The ultrasmple equipment suitable for descriptor systems for dealing with narrow ranges of subject matter seems unlikely to be adequate for the Patent Office requirement.) (iii) training personnel to operate the system.

(2) Cost of operationally establishing and maintaining the system. Here again previously discussed similarities with the uniterm system led to the conclusion that individual patents would have to be reviewed in order to assign an appropriate set of descriptors. As with the uniterm system, it is more than doubtful that the subject headings of subclasses to which patents are assigned could be relied
upon to provide a sufficient set of descriptors for each patent.

The recording of sets of descriptors as assigned to corresponding patents would require a recording medium capable of registering a considerable number of descriptors while avoiding unreliable selection as exemplified by "false drops." For such recording, magnetic tape would appear to be appropriate.

As with the uniterm system, new technological developments and corresponding establishment of new concepts could be expected to lead to extension and revision of the descriptor listing as originally worked out when setting up a descriptor system. Such extension and revision, in turn, would raise the question as to revision of previous analysis of patents embraced by the system.

(3) Cost of operating the documentation system. The similarities with the uniterm system are such that sources of operating costs have much in common as the following resume may make apparent: (a) Costs of deciding what combinations of descriptors corresponds to a given query, i.e., generating a \( \{\text{scdS}\} \triangleleft Q \). (b) Costs of interacting the \( \{\text{scdS}\} \triangleleft Q \) with the \( \{\text{scdS}\} \triangleleft D \), as provided by the system to identify the members of \( \{D\}^n \). (c) Costs of providing to the examiner copies of patents that are members of \( \{D\}^n \). (d) Costs of personal inspection of \( \{D\}^n \) as indicated by the notational expression

\[
\text{psu}(\{D\}^n \rightarrow \{Dp\}^x + \{Dn\}^y).
\]

Previous discussion of these operations in section 4.4 led to the conclusion that a descriptor system when operated to provide a \( \{D\}^n \) that contains all the pertinent patents to a given query is all too likely to require excessive effort to eliminate the unwanted
patents as designated by \( \{Du\}^y \).

5.3.5 ASM-WRU System

(1) Cost of designing the system and providing equipment and personnel. As the demonstration presented in the next chapter of this thesis shows, the design of an information retrieval system to apply the ASM-WRU methods to meet Patent Office search requirements can advantageously make use of the present Patent Office classification both as a source of basic terms and also as a general indication of how the capabilities of the present classification are to be extended. As already discussed in Chapter 4, the ASM-WRU system would provide for each patent a complex symbol as notationally indicated by \( (\{E\}||scmsS) \not\rightarrow D \). In conducting searches, each query is first expressed in a corresponding fashion as \( (\{E\}||scmsS) \not\rightarrow Q \) and automatic electronic equipment is used to detect those patents for which it is the case that \( (\{E\}||scmsS) \not\rightarrow D \rightarrow (\{E\}||scmsS) \not\rightarrow Q \). The patents that meet this requirement are selected for personal review by the examiner.

It is an essential feature of the ASM-WRU system that the generation of the encoded abstracts is accomplished in two steps: (a) personal inspection of the document, in particular a patent, and expression of its important features of subject matter as a standardized abstract. The vocabulary for these abstracts consists of (i) usual English terms as encountered in scientific and technical writing and (ii) role indicators, at least one of which is associated with each term in the standardized abstract. The individual combinations of at least one role indicator plus an English language term are set up in
accordance with analysis of the subject contents of a patent and these combinations are organized into syntactical structures. (b) The individual terms in the standardized abstract are replaced by codes that are worked out so that (i) each term is assigned a unique code (only completely synonymous terms are assigned the same code) and (ii) the code for each term contains one or more code elements (the semantic factors) which specify generic ideas such as amusement (M-SM), device (M-GH), etc. The second letters in the semantic factor codes, here indicated by dashes, designate how the specific term is related to the generic idea as indicated by the semantic factor code element.

As this summary indicates, the costs of designing the system and providing equipment and personnel result from the need to provide the following: (a) A set of instructions for generating the standardized abstracts as referred to above. (See Chapter 5 in Tools for Machine Literature Searching.) (b) Generation of a semantic code dictionary for encoding specific terms as referred to above. (See Chapter 9 in Tools for Machine Literature Searching.) Note in this connection that the encoding of new terms is a continuing operational cost and, as such, may be distinguished from the cost of generating a code dictionary preparatory to initiating operational searching. (c) Definition of procedures for using the semantic code dictionary to convert the specific terms in the standardized abstract into encoded form and thus provide the encoded abstracts for machine searching. (See Chapter 9 in Tools for Machine Literature Searching.) (d) Definition of procedures for encoding queries preparatory to
performing machine searching of the encoded abstract. (See Chapter 15 in *Tools for Machine Literature Searching*.) (e) Definition of procedures for performing the machine searching of the encoded abstracts. (See Chapters 15 and 18 in *Tools for Machine Literature Searching*.)

(f) Training of personnel to perform various tasks as summarized above.

(g) Provision of equipment as needed particularly for (i) encoding the standardized abstracts (see c above) and (ii) for performing searching operations (see e above).

(h) Training examiners how to formulate their queries so that the queries faithfully represent the scope of subject matter of pertinent interest to a given application. Such statement of queries requires, rather obviously, that the claims in the application being examined shall be interpreted in the light of the doctrine of equivalents and related considerations of obviousness to those skilled in the art.

Much of the basic investment in systems development has already been made as far as the cost facts a, b, c, d and e are concerned. The previously developed methods as presented in the cited chapters of *Tools for Machine Literature Searching* could be expected to require adaptation in order to serve Patent Office novelty search requirements to best advantage. Such adaptation would be carried through most efficiently in two steps: (i) Development of a provisional searching system along the lines set forth and demonstrated in Chapter 6 of this thesis. (ii) Test of provisional searching on an operational pilot plant basis to determine what adjustments may be advisable in order to optimize advantages by improving system performance and by minimizing
costs.

With regard to step c (encoding of terms) and e (conducting of searches of encoded abstracts) it may be noted that general purpose digital computers can be and have been programmed to perform these operations. The first development of such programs was carried out by the Computer Department of the General Electric Company for the GE-225 computer.

(2) Cost of operationally establishing and maintaining the system. As far as operationally establishing the system is concerned, the various costs involved have been reviewed and included under the cost of designing the system and providing the equipment and personnel.

In maintaining the system, the following costs must be taken into account: (a) Generating standardized abstracts for newly issued patents. It seems likely that the procedures for preparing standardized abstracts could be worked out so that the examiners responsible for examining a given application could provide the standardized abstract for a patent that matures from an application. Editorial review of encoded abstracts so prepared would be advisable, particularly for less experienced examiners. It should be noted, in this connection, that the major effort in generating a standardized abstract for a document usually is the understanding and analysis of the document’s subject contents. (b) Providing semantic codes for newly encountered scientific and technical terms. The ASM-WRU system provides for encoding of new terms if and when they are encountered. The system is, accordingly, "open-ended" as far as new terminology is concerned.
(c) Encoding the standardized abstracts for newly issued patents and recording the resulting encoded abstracts in a form appropriate for machine searching. As already noted, the operation can be and has been automated almost completely. (d) Training new and replacement personnel in various methods and procedures. (e) Maintaining office facilities and, in particular, the equipment used for encoding standardized abstracts and for conducting searching operations.

(3) Cost of operating the documentation system. In line with the preceding analysis of costs to be anticipated in operationally applying the ASM-WRU methods, the following resume of operating costs round out the cost picture: (a) Costs for deciding how queries are to be encoded preparatory to machine searching, i.e., generating for each query at least one \(\{E\} \sqcup \{scmsS\} \neq Q\). (b) Costs of interacting the \(\{E\} \sqcup \{scmsS\} \neq Q\) with the \(\{\{E\} \sqcup \{scmsS\} \neq D\}\), as provided by the input processing of the system, in order to identify the members of \([D]_n\). It is not necessarily the case that the totality of all previously issued patents would have to be scanned by the searching machine in responding to each query. It is quite feasible to divide up the total multiplicity of encoded abstracts for previously issued patents into general classes in the same way as with the present conventional classification system. (c) Costs of providing to the examiner copies of patents that are members of \([D]_n\). (d) Costs of personal inspection of \([D]_n\) as indicated by the notational expression \(psU([D]_n \rightarrow [Dp]_x + [Du]_y)\).

As with other systems, the usefulness of a system based on the
ASM-WRU methods is to be judged on its overall effectiveness and costs. To be effective for Patent Office novelty search purposes $\{D_p\}^X$ and, by inference, $\{D\}^N$ should contain as members all patents of pertinent interest to a given application. Furthermore, as experience with the conventional Patent Office classification has amply demonstrated, the number of unwanted patents that are provided to the examiner in servicing a query must be kept low. To this end, the ASM-WRU methods, in addition to making it possible to define the scope of searches by combinations of specific terms, also makes available generic terms and syntactical relationships to provide additional discriminating power.

5.4 Conclusion

Cost analysis has been emphasized in this chapter as it must be recognized that an operational system, no matter what its capabilities or elegance of design, must meet the practical requirement of being economically feasible.
CHAPTER 6

SYSTEM DEMONSTRATION

6.1 Introduction

Preceding discussion led to the conclusion that the ASM-WRU system offers promise of successful application to meet the requirements of Patent Office novelty searching. In order to outline the basic steps that would have to be taken to apply this general method, an analysis of one class of the current Patent Office classification will be demonstrated.

Previous review of various documentation systems pointed out that application of the ASM-WRU methods to establish an operational system requires the following: (1) An understanding of performance requirements, that is to say, the purposes that the system must serve. Chapter 1 of this thesis presents a review of novelty search requirements and, in particular, points out the close relationship of such requirements to the claims in patent applications and in patents as issued after examination of the corresponding application. (2) A list of semantic factors and rules for their use in generating codes for specific terms. Before the system can become operational, each document in the file to be searched must be processed to provide an encoded abstract. The abstracts for the totality of documents must be recorded in a medium appropriate for machine searching. Preparatory to conducting a search to respond to a query, the latter is encoded.
In setting up the machine program to conduct the search, specific terms may be designated by their entire codes and generic terms may be designated either by individual semantic factors or by combinations of the same. (3) A list of code indicators and rules for their use in generating encoded abstracts. The role indicators designate the manner of involvement of the various encoded terms in the encoded abstracts as prepared for machine searching. The rules for using role indicators specify how to construct syntactical structures that are the analogs of phrases, sentences, paragraphs and entire articles or messages in literary-style composition.

Thus, the development of an operational form of application of the ASM-WRU methods generates, in effect, a language which enables us to map the subject contents of documents so that selecting and searching operations may be performed by appropriately designed and programmed electronic equipment. It is highly desirable to develop the code language in such a way that: (1) The mapping operation, i.e., the analysis and encoding of documents, may be performed without incurring excessive costs. (2) The costs of converting queries into searching programs and of performing the latter are not excessive. (3) The reliability and selectivity of the system fully meet practical requirements, which must, however, be realistically evaluated if excessive costs are to be avoided. As discussion of the Patent Office novelty search may have made clear, it is scarcely realistic to expect an automated searching system to place on the examiner's desk exactly those documents that he may cite eventually in his decision concerning
the patentable novelty of the subject contents of an application.
This ideal is rendered impractical by the necessity of taking account
of what the prior art would make obvious to those skilled in the art
and, more specifically, the related doctrine of equivalents.

As already stated, the purpose of this chapter is to present
a demonstration of how the ASM-WRU information retrieval methods may
be applied to meet the requirements of Patent Office novelty search-
ing. It was decided to work out this demonstration in a field of mech-
anical invention. For this purpose, Class 46, Amusement Devices, Toys,
was chosen as a mechanical field that embraces a diversity of structur-
al types without encountering unusual problems of complexity as would
be the case in a field such as automatic transmissions for automobiles.

The demonstration itself was worked out along the following
lines: (1) A basic list of terms in the selected field of mechanical
invention was compiled. (2) A test was made to determine how well the
semantic factors previously developed for the ASM-WRU system may be ap-
plied to encode the basic list of terms in the selected field. (3) A
corresponding review of the role indicators which had been previously
developed for the ASM-WRU system was made to consider how well they
may be applied to generate encoded abstracts in the selected field.
(4) Typical patents in the selected field were analyzed as to subject
contents and encoded abstracts were worked out. (5) The capabilities
of a searching system based on encoded abstracts were considered with
respect to the present Patent Office classification system.
6.2 **Compilation of a Basic List of Terms**

The present conventional classification system has been worked out to establish classes and subclasses of maximum usefulness in meeting the requirements of novelty searching. The headings in the conventional classification system define the various classes and subclasses. It appears obvious that the terms in these headings correspond to particularly important features of the subject contents of the patents in the various classes and subclasses.

Accordingly, the basic list of terms was compiled from the subject headings of the conventional classification for Class 46, Amusement Devices, Toys. As a first step toward evaluating the relative importance of the terms in the basic list for working out lists of semantic factors and role indicators, the frequency of occurrence of various terms in the subject head was established by counting.

In making such a count it was necessary to make two decisions. One of these concerned hyphenated words, and it was decided to treat their components as individual words as they usually represent different ideas. This decision does not mean that terms consisting of more than one word would be ruled out of the eventual code dictionary, but the present purpose of arriving at a list of basic terms appeared best served by treating, as individual words, the components of hyphenated words. A second decision was to consider variations of a term as a single term, both for counting purposes and for compiling the basic list of terms. For example, such variations as "containing," "contained," "container," and the like were regarded as variations of the
basic term "contain." Syntactical connective words, especially prepositions, were omitted from consideration in compiling the basic list. Such connectives are not used in ordinary language in such a way that each connective designates some single well-defined relationship. Rather, natural language connectives are used to designate a number of relationships often quite dissimilar. Consequently, the analysis of connectives was not attempted as an aid to developing role indicators.

The above outlined analysis of the subject headings of Class 46, Amusement Devices, Toys generated a rather lengthy list of terms. A partial list of the multiple appearance words is shown in Table I, arranged in order of their frequency of appearance.

It is to this list of words that we look for initial guidance in working out both semantic factors and role indicators which, in phrase structures, may be regarded as serving as basic components in a certain sense, just as semantic factors are the basic components of codes for individual terms.

This list of words provided the starting point for the analysis of this class to determine the components of the retrieval system. In developing the role indicators, the type of relationship portrayed by the higher frequency words is particularly worthy of attention. As indicated by the high frequency word "figure," the form of the object in question is very important. This same idea is also suggested by other words, such as "sheet" or "body." Accordingly, in developing role indicators form or state must be considered. An equally striking feature of these words is the number that portray action,
<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure</td>
<td>20</td>
<td>Toy</td>
<td>13</td>
<td>Move</td>
<td>12</td>
</tr>
<tr>
<td>Operate</td>
<td>10</td>
<td>Means</td>
<td>9</td>
<td>Wheeled</td>
<td>9</td>
</tr>
<tr>
<td>Sound</td>
<td>9</td>
<td>Device</td>
<td>7</td>
<td>Combined</td>
<td>7</td>
</tr>
<tr>
<td>Support</td>
<td>6</td>
<td>Actuated</td>
<td>5</td>
<td>Impeller</td>
<td>5</td>
</tr>
<tr>
<td>Material</td>
<td>5</td>
<td>Motor</td>
<td>5</td>
<td>Self</td>
<td>5</td>
</tr>
<tr>
<td>Element</td>
<td>4</td>
<td>Part</td>
<td>4</td>
<td>Walking</td>
<td>4</td>
</tr>
<tr>
<td>Animal</td>
<td>3</td>
<td>Connector</td>
<td>3</td>
<td>Contain</td>
<td>3</td>
</tr>
<tr>
<td>Cord</td>
<td>3</td>
<td>Projecting</td>
<td>3</td>
<td>Railway</td>
<td>3</td>
</tr>
<tr>
<td>Sheet</td>
<td>3</td>
<td>Trundles</td>
<td>3</td>
<td>Aerial</td>
<td>2</td>
</tr>
<tr>
<td>Air</td>
<td>2</td>
<td>Animated</td>
<td>2</td>
<td>Body</td>
<td>2</td>
</tr>
<tr>
<td>Changeable</td>
<td>2</td>
<td>Coin</td>
<td>2</td>
<td>Dancing</td>
<td>2</td>
</tr>
<tr>
<td>Detonating</td>
<td>2</td>
<td>Electric</td>
<td>2</td>
<td>Handling</td>
<td>2</td>
</tr>
<tr>
<td>Head</td>
<td>2</td>
<td>Hoops</td>
<td>2</td>
<td>Induction</td>
<td>2</td>
</tr>
<tr>
<td>Jet</td>
<td>2</td>
<td>Magnet</td>
<td>2</td>
<td>Mating</td>
<td>2</td>
</tr>
<tr>
<td>Mechanism</td>
<td>2</td>
<td>Multiple</td>
<td>2</td>
<td>Propeller</td>
<td>2</td>
</tr>
<tr>
<td>Rider</td>
<td>2</td>
<td>Rotary</td>
<td>2</td>
<td>Sparking</td>
<td>2</td>
</tr>
<tr>
<td>Spindle</td>
<td>2</td>
<td>Spinning</td>
<td>2</td>
<td>Spring</td>
<td>2</td>
</tr>
<tr>
<td>Spiral</td>
<td>2</td>
<td>Steering</td>
<td>2</td>
<td>Tethered</td>
<td>2</td>
</tr>
<tr>
<td>Top</td>
<td>2</td>
<td>Vehicle</td>
<td>2</td>
<td>Vibrator</td>
<td>2</td>
</tr>
<tr>
<td>Voice</td>
<td>2</td>
<td>Whirl</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE I**

Partial List of Basic Terms Arranged in Order
of Frequency in Subject Headings
such as "move," "operate," "sound," "support," etc. This indicates that action must also be a prime element in the development of role indicators. Further indications as to role indicators are not apparent in this basic word list, but the usefulness of such indications should not be underestimated.

This kind of basic word list (Table I) is of particular importance in working out semantic factors, and it is to this subject that attention will be directed next.

6.3 Semantic Factors and Their Development

Approximately 100 words were selected at random from the basic word list. On checking with the published version of the ASM-WRU dictionary, approximately sixty-five percent of the words were found to be already listed, the remaining words not already listed in the code dictionary were, for the most part, very easily fitted into the existing system by generating codes similar to those shown in Table II. In the codes for the sample words seventy-one of the two hundred and fourteen semantic factors of the ASM-WRU list as published in *Tools for Machine Literature Searching* were used. In other words, these seventy-one semantic factors, at least, appear suitable for metallurgical, chemical, and similar process applications, as well as for the mechanical type of subject matter with which this thesis is concerned. It would be interesting to extend this study to include additional fields within the mechanical arts and also to extend it into the electrical field. It seems likely that additional factors on the ASM-WRU list would prove useful. Furthermore, it seems certain that additional
<table>
<thead>
<tr>
<th>Word</th>
<th>Code</th>
<th>Semantic Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actuated</td>
<td>BACN 009</td>
<td>Origination</td>
</tr>
<tr>
<td>Animal</td>
<td>NAML 001</td>
<td>Animal</td>
</tr>
<tr>
<td>Bodies</td>
<td>BAGP 001</td>
<td>Body</td>
</tr>
<tr>
<td>Combined</td>
<td>GWRP 016</td>
<td>Group</td>
</tr>
<tr>
<td>Connector</td>
<td>GUNT 020</td>
<td>Connection</td>
</tr>
<tr>
<td>Container</td>
<td>CATR 001</td>
<td>Contain</td>
</tr>
<tr>
<td>Detonating</td>
<td>PALS 009</td>
<td>Explosive</td>
</tr>
<tr>
<td>Device</td>
<td>MACH 001</td>
<td>Machine, Device</td>
</tr>
<tr>
<td>Element</td>
<td>LAMN 003</td>
<td>Unit</td>
</tr>
<tr>
<td>Figure</td>
<td>BAPG 015</td>
<td>Body</td>
</tr>
<tr>
<td>Handling</td>
<td>HUND 001</td>
<td>Handling</td>
</tr>
<tr>
<td>Impeller</td>
<td>MACH NURG 007</td>
<td>Device; Energy</td>
</tr>
<tr>
<td>Jet</td>
<td>MAPR MYTN 001</td>
<td>Material property; Motion</td>
</tr>
<tr>
<td>Log Cabin</td>
<td>LAMB BALD 001</td>
<td>Lumber; Building</td>
</tr>
<tr>
<td>Material</td>
<td>MATT 001</td>
<td>Material</td>
</tr>
<tr>
<td>Means</td>
<td>PASS 029</td>
<td>Process</td>
</tr>
<tr>
<td>Motor</td>
<td>MACH NWRG 001</td>
<td>Device; Energy</td>
</tr>
<tr>
<td>Move</td>
<td>MATN 012</td>
<td>Motion</td>
</tr>
<tr>
<td>Part</td>
<td>LYMN RAPR 028</td>
<td>Unit; Relative term</td>
</tr>
<tr>
<td>Railway</td>
<td>HYCL 007</td>
<td>Vehicle</td>
</tr>
<tr>
<td>Rotary</td>
<td>MYTN 019</td>
<td>Motion</td>
</tr>
<tr>
<td>Sparking</td>
<td>MAPR 19%</td>
<td>Material property</td>
</tr>
<tr>
<td>Support</td>
<td>HULP MACH 001</td>
<td>Aid; Machine</td>
</tr>
<tr>
<td>Toy</td>
<td>MUSM 007</td>
<td>Amusement</td>
</tr>
</tbody>
</table>

**TABLE II**
Semantic Codes Generated for Basic Terms
semantic factors will have to be added to the ASM-WRU list in order to provide desired resolving power and to avoid excessive use of certain semantic factors. Three lengthy groups of terms listed under the semantic factors "general property," "physical property," and "material property" are the three prime candidates for subdivision. One factor that certainly needs to be added is "Geometric Shape," whose introduction would considerably reduce the number of terms listed under the semantic factor "General Property."

The methods that were used for generating the codes listed in Table II are described in detail in Chapter 9 of *Tools for Machine Literature Searching*. Each semantic factor is represented by a three letter code, arranged as one letter, a blank and then two letters. The blank is filled by another letter that designates the analytical relationship of the semantic factor to the referent and thus provides discriminating power when using the semantic codes. These letters and their use in generating semantic codes are detailed on pages 278 and 279 of *Tools for Machine Literature Searching* and summarized as listed below.

<table>
<thead>
<tr>
<th>Second Letter Infix</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Categorical</td>
<td>The word coded represents a member of the class represented by the semantic factor.</td>
</tr>
<tr>
<td>E Intrinsic</td>
<td>The word coded represents something composed of that which is represented by the semantic factor.</td>
</tr>
<tr>
<td>I Inclusive</td>
<td>The word coded represents something which is a component of that which is represented by the semantic factor.</td>
</tr>
<tr>
<td>Second Letter</td>
<td>Infix</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>O</td>
<td>Comprehensive</td>
</tr>
<tr>
<td>U</td>
<td>Productive</td>
</tr>
<tr>
<td>Q</td>
<td>Affective</td>
</tr>
<tr>
<td>W</td>
<td>Instrumental</td>
</tr>
<tr>
<td>X</td>
<td>Negative</td>
</tr>
<tr>
<td>Y</td>
<td>Attributive</td>
</tr>
<tr>
<td>Z</td>
<td>Simulative</td>
</tr>
</tbody>
</table>

The above mentioned coding tests with 100 basic terms from the subject headings of Class 46, Amusement Devices, Toys indicated the suitability of the ASM-WRU semantic factors for generating codes, whose usefulness in encoding abstracts of patents will be demonstrated subsequently by preparing encoding abstracts of typical patents. In that connection, the encoding of additional specific terms will also be
demonstrated.

After carrying through these preliminary coding studies, with the 100 basic terms, the remaining words on the list obtained from the subject heads of the classes and subclasses of Class 46 were checked against the ASM-WRU code dictionary. It was found that 74.2% of the words on the list had been encoded already. Most of the remaining 25.8% were encoded with the aid of previously established ASM-WRU semantic factors. Since 3.7% of the words to be encoded portrayed some type of human or animal action, e.g., acrobatics, walking, eating, etc., it is felt that a new correspondingly defined semantic factor should be used in generating codes for the terms in question.

Preparatory to the preliminary studies of ASM-WRU semantic factors for encoding the test list of basic terms, their relationship to various published listings of keywords and descriptors was investigated. It was anticipated that the documentation work of the Armed Services Technical Information Agency (ASTIA), by virtue of its broad interest of a technical nature, might be particularly interesting and helpful. An extensive list of descriptors was obtained from ASTIA, and this list was checked against the 100 test words from the basic list of Patent Office subject heading terms as presented in Table I. The results were disappointing. Less than twenty percent of the 100 test words were directly included among the ASTIA descriptors and even applying broad generalizations only about sixty percent of the 100 Patent Office terms could be adequately covered. The difficulty in expanding the ASTIA descriptor list to include the remaining words
and the loss of specificity in using them confirmed the wisdom of the decision to make use of the ASMW-MLR semantic factor codes.

6.4 Role Indicators

As already noted, analysis of the subject headings for Class 46.9 Amusement Devices, Toys provides only limited guidance in developing role indicators. The reason for this is the extreme terseness of these headings. Accordingly, it was decided to make use of the definitions and explanations of the headings as provided by Classification Bulletin Number 20h. The analysis of these definitions was performed in much the same manner as sentences are diagrammed in basic grammar courses in English. However, instead of indicating nouns, pronouns, verbs, etc., the headings were analyzed so that the terms were diagrammed as to their factual relationships to one another. The development of a list of relationships to be used required considerable thought. The process was to diagram the headings one by one and as appeared necessary, additional alternate statements of the headings were set up.

Examples of such analysis are the following:

1. Class 46-9 Devices utilizing smoke, smoke projecting guns, forming smoke rings, and figure toys in which mechanism is provided to simulate the act of tobacco smoking.

   a. ACTION OBJECT: Device ACTION: Generation OBJECT OF ACTION: Smoke b. ACTION OBJECT: Gun ACTION: Projecting OBJECT OF ACTION: Smoke STATE OF OBJECT OF ACTION: Rings c. ACTION OBJECT: Figure ACTION: Smoking CONDITION ON ACTION: Simulation OBJECT OF ACTION: Tobacco.
2. Class 46-12 Toys related to buildings, lighthouses, fireplaces and parts thereof.
   COMPONENT: Building PROPERTIES OF COMPONENTS: Total or part.

3. Class 46-39 Devices having mechanism in simulation of operating machinery. The devices are not operative structures or models.

4. Class 46-46 Toys related to molds for forming various shapes of plastic material.

A provisional list of relationships that appear highly useful for defining role indicators for this class is as follows:

1. Active object
2. State of active object
3. Action or process
4. Condition typifying or influencing the action
5. Object of action
6. State of the object of the action
7. Components
8. Properties of an object and of a component

This list of relationships was found to be very useful in analyzing the subject headings for Class 252, Chemical Compositions and
proved to be equally applicable. This suggests that they may be effective for a broad range of classes.

6.5 Analysis and Encoding of Patents

The development of semantic factors, their use in codes for terms, and the establishment of role indicators are all preliminaries leading up to the encoding of patents. Example patents for this demonstration were picked at random by the Patent Office from subclasses one through forty-six of Class 46. The ten patents from each subclass as provided by the Patent Office included patents from 1916 through 1961. This part of the investigation was conducted along the following lines: First, detailed review of the specifications of the sample patents to observe what additional characteristics appear most useful as additions to the basic list of terms. Secondly, review of the expanded list of terms as developed by analysis of the headings of the classification system and consideration of their use with the role indicators. Thirdly, review of the expanded list of terms as to development of semantic codes, with special attention to eventual need of additional role indicators. Fourthly, setting up general policies for analysis of patents to generate encoded abstracts. This fourth step is a preliminary to the development of concise and detailed rules for abstracting. Their detailed development is an extensive project in its own right.

In reviewing typical patents, three parts stood out as worthy of special attention: first, the introduction which, in general, outlines the subject matter; secondly, the accompanying diagrams which
are prime aids to understanding the important feature of the subject contents; thirdly, the claims which are, of course, of such importance that they must be adequately abstracted by the encoded abstracts of the retrieval system. This emphasis on claims is justified by well-established procedures in the Patent Office. A patent is assigned to a class on the basis of the prime details in the claims, but it may be cross-referenced to one or more additional classes as may appear advisable in view of important additional features. In marginal cases in which the claims might permit assignment to two different subclasses it appears that the diagrams are used as the deciding factor.

In the preliminary of this part of the investigation, it was easy to discern additional features of subject matter that might be used eventually for expanding the existing classification system. As might be expected, however, these features were not all the same for different subclasses, and in carrying through this approach it would be well to take into account all the subclasses. Likewise there is a multiplicity of additional features that could be used in each case for expanding the classification.

The second part of this phase led to a corresponding conclusion with respect to role indicators. The eight relationships previously mentioned are basically important and can be applied here to indicate outstanding features of subject content. Eventual extension of the basic list of eight role indicators seems likely to prove advisable if broader ranges of subject matter are to be included in an operational searching system.
The third part of this phase produced new words to be added to the list, but here again the ASM-WRU semantic factors were found to be effective for developing codes. It was found possible to characterize the important features of individual terms by applying the semantic codes as explained previously.

The fourth part of this phase is by far the most difficult to work out and to perform. Obviously the key features to be abstracted must come from the claims but the claims are usually much too wordy for encoding purposes. Therefore, when abstracting a patent, it is necessary to be able to recognize the key features and record these while discarding the excess verbage, i.e., excess from the point of view of coding for retrieval purposes. Most patents contain multiple claims and this presents a problem that must be decided by the abstractor. Sometimes the variation from one claim to another is so minor that two or more claims can be well covered by the same coding. However, in other cases the different claims may be sufficiently extensive and the different claims may relate to sufficiently diversified subject matter to require entirely separate coding. Furthermore, in spite of their wordiness, the claims usually do not present more than the most important novel features of the invention. The specification and accompanying drawings must be taken into account in order to attain a sufficiently complete picture of the subject contents of a patent in order to prepare an adequate encoded abstract.

In expressing the important features of subject contents as encoded abstracts, role indicators play a decisively important role.
Accordingly, it seemed advisable to carefully compare the role indicators of the ASM-MRU system with the eight basic relationships that were recognized as important by examination of the classification system headings and their definitions. It seems possible that the usefulness of additional ASM-WRU role indicators might become apparent. The list of role indicators examined was published in *Tools for Machine Literature Searching*.

1. **KOV Property given for** is prefixed to a term that designates a material for which properties are given or implied. Since in our preliminary list properties of objects and of components were recognized as important, this relationship will be included.

2. **KEJ Material processed** is prefixed to a term that designates material that is being acted upon or is being made to perform an action. This corresponds directly with object of action and would, of course, be included.

3. **KUJ Component** is prefixed to a term that designates a constituent of a preceding material or a subdivision that should be isolated for searching. Since this term was on our preliminary list, there is no doubt about its inclusion.

4. **KQJ By means of** is prefixed to a term that designates the actual agent or instrument effecting or used to effect a process, testing technique or function. This relationship fits very well with what was intended as one type of our relationship "condition of action." Although "by means of" is a relationship of lesser scope than "condition of action," its usefulness is evident in light of our preliminary
considerations of encoding patents. "By means of" is accordingly included in our revised set of role indicators.

5. **KWI Product** is prefixed to a term to indicate that it is the result or intermediate result of a process or action. This again fits with the "object of action" relationship or, in some cases, "action object" and therefore will be included.

6. **KAJ Starting material** is prefixed to terms that designate a material which is subjected to processes of fabrication. This again is similar to "action object" and "object of action" and will be included in the revised list of role indicators.

7. **KAD Machine or device** is prefixed to a term designating a machine or device whose description and/or function form the chief subject content or part thereof. This meets the definition of "action object" and, due to the continuous use of the term device in class definitions, must be included.

8. **KAG Subassembly** is prefixed to a term when it designates a subassembly of a previously indicated machine or device. This term is closely related to "component" and on this basis might be questioned. The component of a chemical composition may well be distinguished to advantage from a subassembly of a machine, and to distinguish these two ideas both terms will be included in our list.

9. **KWV Property given** is prefixed to a term that designates some property of a material and/or its components. This role indicator is, obviously, to be included due to its direct tie to "property given for" (number 1 in this list).
10. KUP Property determined is prefixed to a term that designates a property, measurement, value, rate, etc., which has been determined or is the result of a process when the determination or process occurs within the subject content. This role indicator has no more than limited usefulness in an information retrieval searching system for Patent Office needs. It would be useful with patents that relate to scientific instruments or in their applications.

11. KAP Property influenced is prefixed to a term that designates a property stated to be influenced by, dependent upon, variable with or a function of another factor. This role indicator is of principal usefulness in encoding abstracts concerned with such types of processing as heat treating of metals, annealing of glass, etc.

12. KAM Process is prefixed to a term denoting a process, action or function. This role indicator will be included as it specifies a relation previously determined to be important.

13. KXM Process negation is prefixed to a term in order to designate that a process does not occur or is prevented from occurring. This term is, so to speak, the negative counterpart of number 12. This role indicator appears likely to be used rather infrequently for encoding inventions of a mechanical nature.

14. KAH Condition is prefixed to a term to indicate that the term designates ambient or attendant factors or circumstances. The usefulness of this semantic factor was recognized in our preliminary list.

15. KAL Influenced by is prefixed to a term that indicates a
material, process, etc., which influences a preceding property indicated with this role indicator. Comments made concerning "property influenced," number 11 above, apply equally to this role indicator.

16. **KIS Location** is prefixed to a term that designates where a process occurs. This relation is of evident importance in several of the sample patents examined and therefore should be included.

17. **KWB Directed from** or **KWC Directed to** is prefixed to a term that designates the place from which or to which a movement occurs. In more complex mechanical subject matter, role indicators of this type would probably be of importance.

18. **KAB Field** is prefixed to a term that is the name of an industry or scientific field or the name of a theory. This term will not be included as these are not important relations in patent applications.

19. **KIT Time occurring** is prefixed to a term that designates a date as inclusive dates. Since dates are not important features of the subject contents of patents for determining novelty, this role indicator will not be included.

20. **KIG Location-geographic**. Not detailed as, obviously, not applicable.

21. **KIB Organization involved**. Not detailed as, obviously, not of interest in novelty searching of technological subject matter.

22. **KEF Person involved**. Not detailed as, obviously, not applicable.

23. *Namely, with respect to* is used to separate
specifying and descriptive terms from the term specified or described, to separate multiple word terms which can be separated without changing the meaning of the term, to separate a material from a process which is of a prior nature and is of subordinate interest, to separate a material part from a material whole when the whole is of subordinate interest. Although not covered as a relationship in the original list, this general role indicator has obvious applications that should make it of value.

24. NLT Not less than, NMT Not more than, NLMT Exactly.

These role indicators are self-explanatory and since range is an important definition in a patent, it should be included here.

On the basis of this analysis, it has been decided to use numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 16, 23 and 24 as role indicators for encoding Patent Office material. We thus arrive at a total of fifteen role indicators as compared to eight in the original list, but as indicated several relationships are included that were not recognized as important by the preliminary role indicator analysis. The use of "namely, with respect to" should simplify encoding which would otherwise require lengthy combination expressions. The "not less than, not greater than, and exactly" separates an important requirement within the patent law. "Location" better details an occurrence that might be expressed by other means but a gain in precision of expression should also be expected. The use of "component" and "subassembly," and "process" and "process negation" give additional discriminating power although the usefulness of these role indicators did not
become apparent during the preliminary review of sample patents.

6.6 **Analysis and Encoding of Typical Patents**

To demonstrate how the above discussed role indicators may be used with semantic codes of substantive terms to generate encoded abstracts of patents, a number of patents from each of the subclasses available (Class 46, Subclasses 1-46) were encoded. The encoding of three typical patents from subclass 20 is presented here in detail.

In this presentation, first, the claims from the patent are given, secondly, the standardized abstract form of the claims, and then the encoded form of its representation in the system. In the standardized abstract portion the role indicators are set in parenthesis followed by the substantive term, separated by a slash from its semantic factors. In the fully encoded abstract of the patent as presented, commas separate subphrases, dashes separate phrases, and the code elements are as listed in the published ASM-WRU code dictionary.

First example, United States patent number 1,771,783 for a building toy. The claims state:

1. A toy building comprising a plurality of imitation logs formed with angular grooves adjacent their ends, adapted to be assembled in superposed relation with each other to form a predetermined structure, intersecting logs between the adjacent ones of the superposed logs, said superposed logs having interfitting engagement with the intersecting logs as and for the purpose described.

2. In a toy building, a plurality of cylindrical imitation logs of different sizes, each being formed with an angular groove adjacent their ends, and semi-cylindrical logs of different sizes, having semi-angular grooves formed therein for interfitting engagement with the angular grooves formed in the cylindrical logs, when the logs are assembled to form a predetermined structure.
3. In a toy building, a plurality of imitation logs formed with angular grooves adjacent their ends, the angular grooves being formed to receive grooved portions of cooperating logs.


The encoded abstract for this patent: KOV,KAD,LAMB,BADe.001,*_KOV,KAD,MUSM.007_e,KWV,DASM.018.o,KAJ,LAMB,005_e,KAM,CUNV.001_e,KQJ.*_KOV,CATT.008_e,KWV,GAMP.001_e,KIS,TARM.008_e,KIS,LYCN,RAPR.008_e,KOV.*_LAMB.005_e,KWV,GAMP,002_e,KWV,GAMP,003.

Second example, United States patent number 2,059,598 for a toy building construction. The claims state:

1. A toy building equipment comprising in combination a plurality of simulated log members substantially in cross section with notches in their faces adapted to interfit when the logs are arranged in cross relationship so as to permit the body of each log in each wall portion to have flat face engagement with the body of the adjacent log in the same wall portion, and short simulated log members substantially rectangular in cross section having notches in substantially centered position longitudinally thereof and adapted to have interfitting engagement with ends of the first mentioned logs at a wall opening such as a doorway, certain of said short log members being of half the thickness of said first mentioned square log members with a notch in only one face and being adapted by flat face engagement with the flat face portion of the bodies of adjacent logs to reinforce and strengthen the building structure.
2. In a toy building construction, the combination of a plurality of simulated log members notched at their ends and built up in crossed interfitting relationship to form the body of a toy cabin, triangular gable end members at opposite ends of the cabin having registering openings therethrough, a horizontal pin mounted in said openings and held by friction therein for holding the triangular members in upright position, and simulated plank pieces mounted on said triangular members providing a roof enclosure for the cabin.

3. In a toy building construction, the combination of a plurality of simulated log members notched at their ends and built up in crossed interfitting relationship to form the body of a toy cabin, triangular gable and members at opposite ends of the cabin having registering openings there-through, a horizontal pin mounted in said openings and held by friction therein for holding the triangular members in upright position, upwardly extended pins in said triangular members, and simulated plank pieces mounted on said triangular members and releasably engaging said upwardly extending pins providing a roof enclosure for the cabin.

4. In a toy building construction, the combination of a plurality of simulated log members notched at their ends and adapted to be built up in crossed interfitting relationship to form the body of a toy cabin, two triangular gable end members arranged end to end at opposite sides of the cabin at each end, horizontal pins connecting the triangular members at each end of the cabin for holding them in operative position, and simulated plank pieces mounted on said triangular members providing a roof enclosure for the cabin.

Standardized abstract for encoding this patent: 1. (Device) (Property given for) Log cabin/lumber-building (Namely) Toy/Amusement (Property given) Knocked down/dismantled. 2. (Starting material) (Logs/lumber (Process) Connected/connect (By means of) (Property given for) Notches/cut (Property given) Rectangular/geometric shape (Location) End/end (With respect to) Adjacent/location-relative term (Location) Center/location-material property. 3. (Property given for) Logs/lumber (Property given) Rectangular/geometric shape. 4. (Starting material) (Property given for) Gable/building-cover (Property
In a toy building construction, the combination of a plurality of logs in crossed relationship to each other and having notches in their faces at a short distance from their ends for providing interlocking engagement with each other, gable members at oppositely sloping end faces and extending at their ends beyond the outer faces of the side walls of the building, roofing boards mounted on said gable members at opposite sides of the building, and clips formed of metal strips bent into the form of sockets adapted to have a strong grip on the upper logs of the side walls of the buildings and extending sidewise from the sockets and bent to provide hooks at a substantial distance outwardly beyond the ends of the gable members so as to engage the lower edges of said roofing boards for holding said boards in position on said gable members.
fabricated. 4. (Starting material) Plank/lumber (Process) Fastened/
fasten (By means of) Hooks/device (Product) Roof/building-cover.

The encoded abstract for this patent: KOV.KAD.LAMB.BALD.001.<,
KOV.KAD.MUSM.007.<,KWV.DASM.018.<,KAJ.LAMB.005.<,KAM.CUNT.001.<,KQJ.CATT.
008.<,KIS.TARM.008.<,KIS.LYCN.RAPR.008.<,KOV.BUILD.CACR.001.<,KWV.CANS.
013.<,KAJ.LAMB.006.<,KAM.FAST.001.<,KQJ.FUST.MACH.021.<,KWJ.BYLD.CACR.002.<

6.7 Searching of Encoded Abstracts of Patents

As already pointed out in summarizing the ASM-WRU information
retrieval methods, the performance of a search to respond to a given
query may be regarded as consisting of three steps: (1) Statement of
query in the form of standardized phrasing which is consistent with the
standardized abstracts of the documents to be searched. (2) Conversion
of the standardized phrasing of the query into encoded form which
is consistent with the encoded form of the standardized abstracts of
the documents to be searched. (3) Interacting the encoded form of the
query with the encoded abstracts of the documents to be searched to
identify those documents whose encoded abstracts contain as a subset
the specific combination of elements of the encoded query. This final
step is particularly simple and easy to demonstrate when the specific
combination of elements of the encoded query involves only the rela-
tionship of logical product among the elements, as in the following
examples:

First search question: Disassemblable toy log cabins. This
question in standardized form is: (Device)(Property given for) Log
cabin/lumber-building (Namely) Toy/amusement (Property given)
Disassemblable/dismantle. The encoded form of this question is: KOV.

A search on this question would produce all three patents encoded, as all the elements in the encoded form of the question are included in the first phrase of the encoded abstracts for each of the patents. It should be noted that this question is equivalent to and would group these patents together as in the existing classification system.

Second search question: Wooden buildings. This question in standardized form is: (Device) Wooden building/lumber-building. The encoded form of this question is: KAD.LAMB.Bald.

Here again all three patents would be selected as the elements in this encoded query are all contained in the first phrase of the encoded abstract for each of the patents.

Third search question: Knock-down toy log cabin, with prefabricated gables and plank roof. This question in standardized form is: 1. (Device) (Property given for) Log cabin/lumber-building (Namely) Toy/amusement (Property given) Knock-down/dismantle 2. (Property given for) Gables/building-cover (Property given) Prefabricated/fabricated. 3. (Starting material) Plank/lumber (Process) Connected/connect (Product) Roof/building-cover. The encoded form of this question consists of three phrases: KOV.KAD.LAMB.Bald.001.R.KOV.KAD.MUSM.007.o.KWV.DASM.

This search would select only the latter two patents but not
the first one, as the element combination specified by the search is found in the encoded abstracts for the latter two patents only. Specifically, the elements of the three phrases in the code combination for the query are found in phrases 1, 4 and 5 of the encoded abstract for the second patent and in phrases 1, 3 and 4 of the encoded abstract for the third patent.

Fourth search question: Knock-down toy log cabin with rectangular shaped logs and connected by rectangular notches. The standardized form of this question is: 1. (Device) (Property given for) Log cabin/lumber-building (Namely) Toy/amusement (Property given) Knock-down/dismantle. 2. (Property given for) Logs/lumber (Property given) Rectangular/geometric shape. 3. (Starting material) Logs/lumber (Process) Connected/connect (By means of) (Property given for) Notches/cut (Property given) Rectangular/geometric shape. The encoded form of the question also consists of three phrases, namely: KOV.KAD.LAMB. BALD.001. KOV.KAD.MUSM.007.KWV.DASM.018. KOV.LAMB.005.KWV.GAMP.003. KAJ.LAMB.005.KAM.CUNT.001.KQJ.KOV.CATT.008.KWV.GAMP.003.

This search would select only the second patent shown above, as the elements of the three phrases of the code combination for this search question is found in phrases 1, 3 and 4 of the encoded abstract for the second patent. Such inclusion is not observed with the encoded abstracts of the other two patents.

These example searches, as presented here, are relatively simple and are not intended to show how the semantic code and role indicators may be exploited in order to provide maximum discriminating
power. In particular, the example searches do not make use of the logical sum and logical difference for encoding of the questions to be searched. For details in this connection, see Chapter 15, *Tools for Machine Literature Searching*.

As may be evident from these example searches many other combinations of the code elements in this type of encoded abstracts could be set up to define searches. The result of each such search would be a set of patents. The number of such sets that could be established would be equal to the number of possible combinations of code elements and this number is, of course, very large. The overall result is, accordingly, the effective extension of the present conventional classification system by making it possible to generate a specially defined subclass when and as it may be useful for novelty search purposes.

6.8 Conclusion

The demonstration presented in this thesis led to the following conclusions concerning the application of the ASM-MRU methods to the novelty search problem of the Patent Office: (1) Searches of encoded abstracts of patents may be performed by automatic devices to select sets that correspond in generic scope to the classes and subclasses of the present classification system. (2) The generic-specific scope of the sets of patents selected by machine searching of encoded abstracts may be varied within very wide limits even though the encoding system is relatively simple. (3) The conventional classification system of the Patent Office can provide both the basis and also highly
useful guidance for developing methods for generating encoded abstracts of patents.
REFERENCES


4. Patent Office Research and Development Reports:


