CLM - R101

United Kingdom Atomic Energy Authority
RESEARCH GROUP
Report

THE STATUS PROJECT :
SEARCHING ATOMIC ENERGY LAW BY COMPUTER

G. B. F. NIBLETT
N. H. PRICE

Culham Laboratory
Abingdon Berkshire
1969
C.13
Available from H. M. Stationery Office
THREE SHILLINGS NET

U.D.C.
681.3:025.3

CLM-R 101

THE STATUS PROJECT : SEARCHING ATOMIC ENERGY LAW BY COMPUTER

G.B.F. Niblett and N.H. Price

ABSTRACT

A suite of computer programs has been prepared for mechanised searching of Acts of Parliament. Application of these programs to the full text of statutes and statutory instruments dealing with atomic energy is described. A detailed account of the search language is presented together with typical examples of the searching process. Further development of these information retrieval techniques is discussed.

UKAEA Research Group
Culham Laboratory
Abingdon
Berkshire.

November, 1969

CONTENTS

1. Introduction

The last few years have seen a rapid growth in the use of digital computers for mechanised processing of documents and this subject promises to be one of the most fruitful areas of computer application. Interest in the automatic handling of textual matter is prompted at least in part by the remarkable developments in speed and capacity of computing systems : computers can now process words at speeds approaching a million a second, and storage devices are in use which provide rapid access to up to 10^{10} words. The predicted growth of computer typesetting techniques, and the increasing reliability of devices for optical character recognition makes it likely that in the next couple of decades most important documents and much of the world's literature will be available in digital form. As a consequence the full range of computer facilities becomes available for processing these texts : they can be amended and up-dated by computer; dictionaries, concordances and indexes can be automatically generated; the text can be rapidly and accurately searched by computer; and perhaps most valuable of all, the logical properties of the computer can be used to detect and possibly remove syntactic ambiguities. It seems likely therefore that in the next few years the computer will come to play an increasing and possibly paramount role in the preparation and communication of the written word.

A particularly interesting application of these techniques is to legal documents such as statutes and statutory instruments. Such documents are clearly important and a rapid and efficient retrieval service is undoubtedly needed. The formal and precise language used in statutes offers the hope that such documents can be effectively searched and syntactically analysed by computer. The case for applying computer techniques to U.K. Acts of Parliament and subordinate legislation is argued in Ref.(1) which gives an idea of the magnitude of the task and the likely advantages. The text of the statute book is continually being modified by repeal, amendment and consolidation and the use of the computer would enable the full up-dated text (if necessary with annotations) to be maintained in a form suitable for immediate publication as an authorised version by computer typesetting methods.

The suitability of legal documents for storage and search by computer is exemplified by the current interest being shown in legal information retrieval. A pamphlet published in 1966[2] reported that in the U.S.A. alone there were more than thirty research projects, and there are at least two commercial services for mechanised law searching presently in operation. A recent review paper by Fraenkel[3] provides a valuable discussion and appraisal of this current work. It is worth noting that there is a sense in which information retrieval has a greater directness and immediacy when applied to the law than when applied to other subjects. In, say, physics or the social sciences what is retrieved is information <u>about</u> these subjects : but in the case of the law what can be retrieved are the direct words of the law <u>itself</u>, insofar as the law is comprised in the text of Acts of Parliament and subordinate legislation.

Legal information retrieval projects fall naturally into two categories : those which store the full text, or the full text other than common words such as 'the', 'of', 'or', 'and', etc. ; and those which store an abstract or digest of the text. In view of the rapid growth in computer storage systems it appears more satisfactory to search and analyse the full text; this avoids reliance on the fallible efforts of human abstracters

who cannot possibly foresee all future applications and implications of the material they are abstracting. Full-text searching is especially desirable for statute law where the precise words are so important - no digest or abstract can possibly be adequate.

Two particularly interesting applications of the computer to full-text searching are those of Horty[4][5] in the U.S.A. and Tapper[6][7] in the United Kingdom. Horty's work began in 1959 and has since developed to the point where his group has available for searching the full text of the statutes of all the States and the Federal legislature. Horty's team was originally part of the University of Pittsburgh but has since been transferred to the Aspen System Corporation of which the University is a principal shareholder. This company operates a commercial retrieval service using IBM document processing programs. It has been reported[8] that investment in the programs represents more than fifty man-years and a cost of some 500,000 U.S. dollars. Colin Tapper's research at Oxford university is noteworthy for two reasons. Firstly, he has pioneered the application of computer searching techniques to the full text of case law; and secondly, his preliminary results, reported at the 1967 World Peace Through Law Conference in Geneva, show that in terms of finding relevant documents computer searching compares favourably with searching by conventional methods.

In view of the likely advantages in terms of speed and accuracy of processing documents by computer, a suite of programs has been developed on the KDF9 computer at the Culham Laboratory suitable for serving the internal purposes of the Atomic Energy Authority. These programs have been applied to the full text of U.K. statutes and statutory instruments dealing with atomic energy, a text which currently amounts to some 138,000 words.

The purpose of this study (known as the STATUS project from STATUte Search) can be briefly stated : to develop computer software suitable for organising the storage, indexing, analysis and searching of the full text of lengthy documents. The programs have the following objectives:-

(a) to process the full text of documents rather than abstracts or summaries of the text;

(b) to be written primarily in a high-level language available on a wide variety of computers so that the programs can be to some extent machine independent;

(c) to incorporate a search language (called QUEST for QUEry STatute) designed for interrogation of the documents;

(d) to incorporate programs suitable for searching both off-line and on-line, i.e. in batch-mode and conversational mode using teletype consoles.

The present paper is a progress report on the STATUS project and describes results obtained at the completion of the first phase of the exercise.

2. Atomic Energy Legislation

There are a number of reasons why U.K. atomic energy legislation is suitable for a small-scale exercise in statutory information retrieval. Firstly, the size of the text is such that it can readily be handled on the KDF9 machine.

Secondly, the text consists of both statutes and statutory instruments and so can provide experience in handling both types of document. Thirdly, the legislation is relatively self-contained: there is little U.K. case law relating directly to atomic energy so that virtually all the technical law is comprised within the small compass of the statutes and statutory instruments. Finally, though the legislation is all of recent origin, its context is heterogeneous and diverse. In Street and Frame's words[9], "the law relating to nuclear energy is of outstanding interest to lawyers because of the original solution it contains to the new issues presented". The atomic energy legislation creates new criminal offences; it deals with financial matters relating to atomic energy; it creates a statutory corporation, the Atomic Energy Authority; it imposes on that Authority and other operators of nuclear installations a liability more strict than that imposed by any other legislation; in brief, though consisting of only 138,000 words, it embraces a wide variety of subject matter suitable for computer searching.

When storing Acts of Parliament in a computer it is convenient to choose as the unit of text the section or schedule rather than an Act as a whole. Each section is therefore taken to be a document for searching purposes. Apart from the natural convenience of this procedure there is a sound legal reason for it since each section of an Act is a substantive enactment.* In the case of statutory instruments it is the article or regulation that is taken to be the document for searching purposes.

The size of the atomic energy legislation currently stored on magnetic tape is shown in Fig.1. It consists of 161 documents drawn from Acts of Parliament and 610 documents from Statutory Instruments. The precise number of words is 138,661. The earliest statute is the Atomic Energy Act 1946 and the most recent the Nuclear Installations Act 1969. The text is constantly changing as new Acts or Instruments are added and older ones amended, repealed or revoked.

Statistics of U.K. Atomic Energy Legislation Currently in Force

	No. of documents$^{\phi}$	No.of sentences	No.of words
Acts of Parliament	161	843	49,978
Statutory Instruments	610	1,247	88,683
Total	771	2,090	138,661

$^{\phi}$ A document is defined as a section or schedule of an Act, or an article or regulation of an Instrument.

Fig. 1

*Before 1850 it was the practice to preface each separate portion of an Act with enacting words; since the coming into force of Lord Brougham's Act 1850 this has no longer been necessary.

3. Digital Conversion of the Text

The first step in computer analysis of documents is to convert the text into digital form so that it can be stored on magnetic tape and processed by the computer. There are a number of ways in which this can be done: for example, by the use of punched cards, or paper tape, directly encoding the text on to magnetic tape or using optical character reading systems. The optical methods appear the cheapest and most promising in the long term and preliminary experiments on reading directly on to tape from typewritten copies of the Act are proceeding. However it seemed simplest for the early work to prepare the magnetic tape from punched cards using upper case only and this is what has been done.

The text has been punched onto cards starting in column 1 and using one card for each line of the printed text. Examples of the text as it appears when listed by the computer are shown in Figs.2 and 3. Each line is punched exactly as it appears except for the following changes: upper case only is used; each line is started in column 1; the inverted commas ' and " are interpreted by an equals sign; the symbols .. and ., are used to represent the colon and semicolon; and a hyphen to denote a broken word at the end of a line is replaced by a plus sign. In addition the title of an Act of Parliament which is cited in the text is preceded and followed by an asterisk. The strings of characters between asterisks are treated by the computer as single words : this is a convenient way of tracing statutory citations so that the computer can treat them in a special manner.

Experience shows that no more than 76 columns of the cards are required to contain each line of text. Columns 78, 79 and 80 are ignored by the program and can be used to number the cards. Column 77 is used to store an alphabetic character, termed the card code, which specifies the nature of the text on that particular card, for example whether the text is part of a long title, or schedule, or marginal note. Examples of code letters are as follows :-

Code letter	Nature of corresponding text
A	annotation to an Act or Instrument
E	explanatory note to an Instrument
I	heading to a section or sections
J	title of document
K	long title of an Act
L	text of section of an Act
M	marginal note
O	text of Statutory Instrument
S	text of Schedule to an Act

All the words of an Act or Instrument are punched on cards. Each document begins with a card containing a title which uniquely describes it; this card is given the code letter J. Similarly, all marginal notes are included with code letter M. If the text of a document has been amended or in part repealed this is noted in the document and the annotation given the code letter A. When the words are later concorded the computer is instructed to ignore J and A cards; words on those cards are not part of the text of

ATOMIC ENERGY AUTHORITY ACT 1954 CH 32 SEC 1. J718

THE UNITED KINGDOM ATOMIC ENERGY AUTHORITY. M717

(1) THERE SHALL BE AN AUTHORITY, TO BE CALLED THE UNITED L719
KINGDOM ATOMIC ENERGY AUTHORITY (HEREAFTER IN THIS ACT REFERRED L720
TO AS =THE AUTHORITY=), WHO SHALL, AS FROM THE APPOINTED DAY, L721
EXERCISE AND PERFORM THE FUNCTIONS ASSIGNED TO THEM BY THIS ACT. L722
(2) THE AUTHORITY SHALL CONSIST OF A CHAIRMAN AND NOT LESS L723
THAN SEVEN NOR MORE THAN FIFTEEN OTHER MEMBERS. L724
(THIS SUBSECTION HAS BEEN AMENDED BY SECTION 1 OF THE ATOMIC A724
ENERGY AUTHORITY ACT 1959) A725
(3) ALL THE MEMBERS OF THE AUTHORITY SHALL BE APPOINTED BY L725
THE LORD PRESIDENT OF THE COUNCIL AND OF THOSE MEMBERS- L726
(A) THREE SHALL BE APPOINTED FROM AMONGST PERSONS APPEARING L727
TO THE LORD PRESIDENT OF THE COUNCIL TO BE PERSONS WHO L728
HAVE HAD WIDE EXPERIENCE OF, AND SHOWN CAPACITY IN L729
DEALING WITH, PROBLEMS ASSOCIATED WITH ATOMIC ENERGY., L730
AND L731
(B) ONE SHALL BE APPOINTED FROM AMONGST PERSONS APPEARING L732
TO THE LORD PRESIDENT OF THE COUNCIL TO HAVE HAD WIDE L733
EXPERIENCE OF, AND SHOWN CAPACITY IN, ADMINISTRATION L734
AND FINANCE., AND L735
(C) ONE SHALL BE APPOINTED FROM AMONGST PERSONS APPEARING L736
TO THE LORD PRESIDENT OF THE COUNCIL TO HAVE HAD WIDE L737
EXPERIENCE OF, AND SHOWN CAPACITY IN, THE ORGANISATION L738
OF WORKERS. L739

Fig.2

ATOMIC ENERGY AUTHORITY ACT 1954 CH 32 SEC 2. J783

FUNCTIONS OF THE AUTHORITY. M782

(1) ON THE APPOINTED DAY, THE AUTHORITY SHALL TAKE OVER L784
FROM THE LORD PRESIDENT OF THE COUNCIL THE CARRYING ON OF THE L785
ACTIVITIES THEN BEING CARRIED ON BY HIM UNDER SUBSECTION (1) OF L786
SECTION TWO OF THE *ATOMIC ENERGY ACT 1946*, AND SUBSECTION(1) OF L787
SECTION ONE OF THE *RADIOACTIVE SUBSTANCES ACT 1948*, AND THE PROVI+ L788
SIONS OF THE SECOND SCHEDULE TO THIS ACT SHALL HAVE EFFECT IN RELATION L789
TO THE PROPERTY, RIGHTS AND LIABILITIES HELD OR ENJOYED BY, OR L790
INCUMBENT ON, THE LORD PRESIDENT OF THE COUNCIL FOR THE PURPOSES L791
OF OR IN CONNECTION WITH THOSE ACTIVITIES. L792
(2) SUBJECT TO THE PROVISIONS OF THIS ACT, THE AUTHORITY SHALL, L793
AS FROM THE APPOINTED DAY, HAVE POWER (WHETHER WITHIN THE L794
UNITED KINGDOM OR ELSEWHERE)- L795
(A) TO PRODUCE, USE AND DISPOSE OF ATOMIC ENERGY AND CARRY L796
OUT RESEARCH INTO ANY MATTERS CONNECTED THEREWITH., L797
(B) TO MANUFACTURE OR OTHERWISE PRODUCE, BUY OR OTHERWISE L798
ACQUIRE, STORE AND TRANSPORT ANY ARTICLES WHICH IN THE L799
OPINION OF THE AUTHORITY ARE, OR ARE LIKELY TO BE, REQUIRED L800
FOR OR IN CONNECTION WITH THE PRODUCTION OR USE OF L801
ATOMIC ENERGY OR SUCH RESEARCH AS AFORESAID, AND TO L802
DISPOSE OF ANY ARTICLES MANUFACTURED, PRODUCED, BOUGHT L803
OR ACQUIRED BY THEM., L804
(C) TO MANUFACTURE OR OTHERWISE PRODUCE, BUY OR OTHERWISE L805
ACQUIRE, TREAT, STORE, TRANSPORT AND DISPOSE OF ANY L806
RADIOACTIVE SUBSTANCES., L807
(D) TO DO ALL SUCH THINGS (INCLUDING THE ERECTION OF BUILDINGS, L808
AND THE EXECUTION OF WORKS AND THE SEARCHING FOR AND L809
WORKING OF MINERALS) AS APPEAR TO THE AUTHORITY L810
NECESSARY OR EXPEDIENT FOR THE EXERCISE OF THE FOREGOING L811
POWERS., L812

Fig.3

COMPUTER PROGRAM FOR ANALYSING AND SEARCHING FULL TEXT OF STATUTES

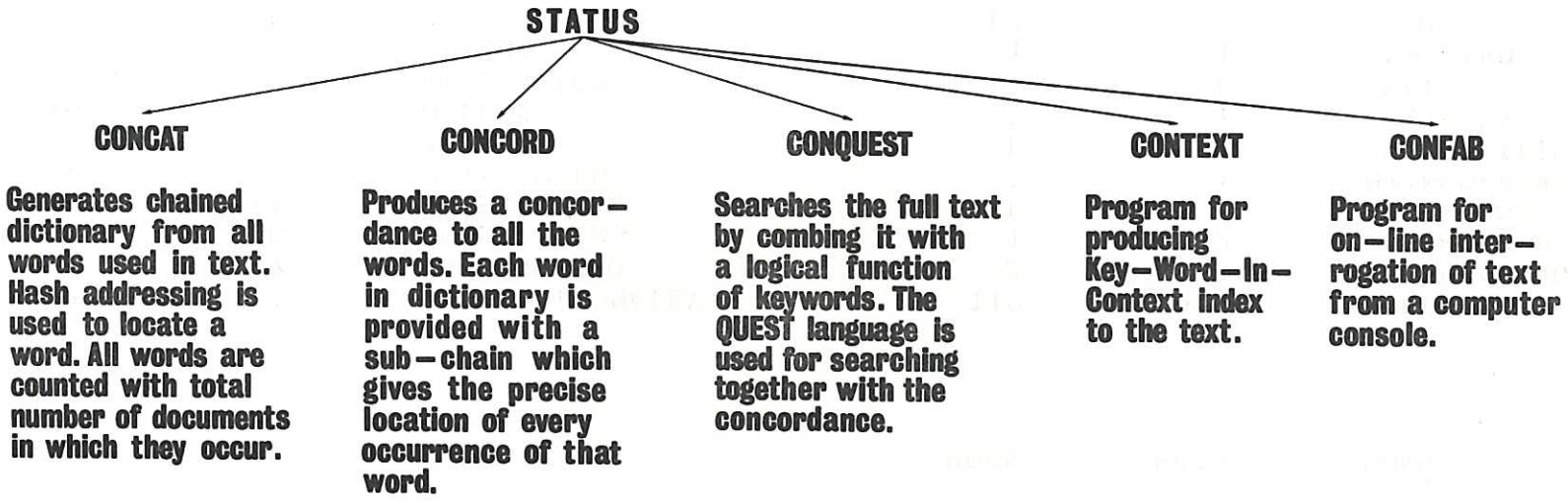

Fig.4

DICTIONARY

DOCS	FREQ	WORD	DOCS	FREQ	WORD
U					
2	2	UNCONDITIONALLY	110	412	UNDER
6	7	UNDERGROUND	2	2	UNDERTAKINGS
12	20	UNDERTAKING	1	2	UNDERTAKEN
7	10	UNDERTAKERS	1	1	UNDERTOOK
1	1	UNINCLOSED	1	1	UNINCORPORATE
28	66	UNITED	1	2	UNIVERSITIES
26	31	UNLESS	1	1	UNLIKELY
2	4	UNOCCUPIED	10	11	UNTIL
1	1	UNTRUE	1	1	UNWANTED
7	8	UP	12	22	UPON
5	12	URANIUM	1	1	URBAN
1	1	URGENCY	33	70	USE
22	56	USED	2	2	USERS
2	2	USES	4	4	USING
2	2	USUAL	2	2	UTILISING

Fig.5

the Act and are therefore not included in the dictionary or concordance.

Fig.2 shows part of the Atomic Energy Act 1954 s.1 as listed by the computer. Subsection 2 of this section has been amended, and this amendment has been incorporated in the text and noted on the A cards. In this way the computer stores an up-dated and annotated version of the legislation. Examples of asterisk strings can be seen in section 2 of the 1954 Act (Fig.3) where reference is made to the Atomic Energy Act 1946 and the Radioactive Substances Act 1948.

4. The Computer Programs

The general design of the STATUS suite of computer programs is summarised in Fig.4. All five sub-programs have been written primarily in a high-level language in order that changes can readily be made and to simplify conversion to another computer. With the exception of CONFAB (written mainly in ALGOL) the programs are in FORTRAN with the addition of short user-code sub-routines for those parts of the program (dealing with character-shifting for example) which are used frequently and which are best written in a low-level language.

The programs have not yet reached their final form since improvements are regularly made in response to new suggestions and ideas. It seems best to describe the content and function of the programs by considering each in turn.

(a) CONCAT. This program is designed to generate a chained dictionary of words in the text together with their frequency of occurrence and the number of documents in which they appear. Words of length up to and including 90 characters are permitted. Every word is included in the dictionary but figures and punctuation signs are excluded. Asterisk strings are treated as special words which come alphabetically after the letter Z. Thus titles of Acts of Parliament occur together in alphabetical order at the end of the dictionary; this is a convenient way of grouping those citations together for rapid reference.

Fig.5 shows those words in the atomic energy Acts of Parliament which begin with the letter U, whilst the thirty most frequently occurring words are listed in Fig.6. The word 'the' amounts to some 8% of the text and the six words 'the', 'of', 'or', 'to', 'in', and 'and' together add up to more than one-quarter of the text.

List of Most Frequent Words

THE	11133	THAT	1762	NOT	930
OF	9719	THIS	1680	ON	887
OR	4259	WHICH	1642	FROM	861
TO	4179	FOR	1600	ORDER	826
IN	4110	BE	1566	IT	824
AND	3216	AS	1496	SUCH	786
A	2603	SHALL	1266	AN	768
ANY	2369	RADIOACTIVE	1104	WASTE	762
IS	1982	ACT	1027	ARTICLE	746
BY	1764	SECTION	1014	UNDER	711

Number of different words = 3,807
Total number of words = 138,661

Fig.6

CONCORDANCE

DOCS	FREQ	WORD					
7	10	TAKE					
		48. 2. 8.L,	5 . 6.151.L,	57. 7. 23.S,	57. 9. 14.S,	81. 3.186.L,	81. 5. 7.L,
		84. 3. 74.L,	112. 7.104.L,	155. 5.131.S,	155. 8.141.S,		
16	29	TAKEN					
		8. 3. 55.L,	10. 4. 50.L,	16. 3. 13.L,	23. 5. 92.S,	30. 2. 43.L,	51. 6.156.L,
		62. 3. 20.L,	62. 5. 83.L,	62. 5. 96.L,	62. 5.161.L,	81. 2. 58.L,	81. 4. 73.L
		81. 5. 34.6,	81. 5. 91.L,	91. 3. 24.L,	101. 7. 99.L,	111. 6. 23.L,	112. 2. 13.L,
		112. 3. 14.L,	112. 4. 14.L,	112. 7. 19.L,	119. 1. 29.S,	119. 6. 15.S,	119.12. 54.S,
		119.13. 94.S,	119.13.121.S,	128. 2.143.L,	137. 2. 60.L,	140. 3. 44.L,	
2	2	TAKES					
		64. 7. 97.L,	139. 3. 57.L,				
5	5	TAKING					
		51. 3.107.L,	81. 5.114.L,	98.10. 16.S,	119. 4. 56.S,	134. 3. 12.L,	
3	14	TAX					
		52. 3. 7.L,	52. 3. 22.L,	52. 3. 47.L,	52. 4. 1.L,	52. 4. 57.L,	52. 4. 68.L,
		52. 4. 72.L,	105. 4. 10.S,	105. 4. 80.S,	123. 2. 12.S,	123. 2. 28.S,	123. 2. 63.S,
		123. 2. 74.S,	123. 2. 78.S,				
1	1	TAXES					
		23.11. 41.S,					
1	1	TAXING					
		151. 2.271.L,					
1	2	TEACHING					
		86. 2. 37.L,	86. 2. 59.L,				
1	1	TECHNICAL					
		114. 4. 27.L,					
9	16	TECHNOLOGY					
		109. 1. 30.K,	110. 3. 56.L,	110. 3. 66.L,	113. 2. 38.L,	114. 1. 10.M,	114. 2. 9.L,
		114. 3. 4.L,	115. 2. 21.L,	118.15. 29.S,	118.15. 61.S,	118.16. 25.S,	118.17. 26.S,
		118.17. 54.S,	140. 4.141.L,	145. 5. 72.L,	147. 3. 51.L,		

Fig.7

(b) CONCORD. This program which shares many common sub-routines with CONCAT is designed to produce a concordance or inverted index to the text. By concordance is meant a listing of each word giving the precise location of each occurrence of that word. The location is specified by the document number, the sentence number within the document, the position of the word in the sentence and the code value of that line of text. This information about each text word is packed away in one 48-bit computer word : up to 262,144 documents are possible with up to 4096 sentences in a document and 4096 words in a sentence. The locations of each dictionary word are chained together in order of occurrence and linked to the dictionary entry.

The CONCORD program provides the option of specifying any word as a common word. A common word is not concorded and thus this facility is able to save storage space. Some words, for example the six most common words, may be of little use in searching, and therefore it is convenient to exclude them from the concordance. In any case, when the concordance is printed it is helpful to exclude listing the location of common words. In practice for the off-line search program all words have been concorded, whereas for on-line operation for which disc storage space is at a premium, about 100 common words have been excluded. This reduces the size of the disc-stored concordance to less than half.

Part of the print-out of the concordance to the atomic energy Acts is shown in Fig.7. The word 'technology', for example, occurs sixteen times in nine documents starting with document 109 and ending with document 147. By inspection of the code values it is seen that the word occurs in a long title, in a marginal note, in a schedule, as well as in various sections of Acts of Parliament.

(c) CONQUEST. Computer searching of the text off-line, that is using batch processing methods, is the function of this program. The text is combed for the presence or absence of selected words with a prescribed proximity to each other; in practice it is the concordance that is searched rather than the text so that it is necessary to examine only those words that appear in the search enquiry. The program can handle many separate searches on one computer run - the limit is set by the size of core storage available. Searching atomic energy legislation on the KDF9 computer, up to fifty searches have been carried through on one computer run.

Search requests are framed in terms of the QUEST language. The result of a search is the number of documents which satisfy the search, the titles (i.e. the J cards) of the documents and if required, a print-out of the text of the documents. It is normal to specify an upper limit on the number of titles and documents to be printed so as to avoid printing a large amount of material unnecessarily. Usually all that is required is the title of the document.

(d) CONTEXT. The program is designed to provide a Key-Word-In-Context (or permuted) index to the text of the form first proposed by Luhn[10]. It has been written by Miss N.E. Couchman and takes as the starting-point for generating the index the concordance provided by CONCORD. The program allows various standard options: for example, specified words can be excluded from the list of words to be indexed and also from the context provided. The format can be varied: the number of columns per page, one or two can be specified, as can the number of contextual characters on either side of the keyword. The program is written in FORTRAN and therefore has the advantage that modifications and

improvements can readily be incorporated.

(e) CONFAB. This program organises the on-line interrogation of text from a computer console using the QUEST language in interactive mode. If a search enquiry yields unsatisfactory results, too few or too many documents retrieved for example, the search enquiry can be rephrased at the console. This type of searching therefore approximates more closely to the conventional type of manual search in which a searcher leafing through an index, or the pages of a book, can continuously widen or narrow the scope of his enquiry. For example, if a poorly phrased enquiry leads to only one or two documents being recovered when it is expected that many more are relevant, the language of the retrieval document can itself be incorporated in the next search to refine the process. The CONFAB program has been written by Dr.M.D.Poole and a comprehensive account of its structure and mode of operation is in course of preparation.

5. The QUEST Language

QUEST is a simple interrogation language for specifying the requirements of any search query and consists of rules for defining the relationship between selected keywords. Two types of relationship are used: logical and positional. The logical relationships provide rules for the absence or presence of words; the positional relationships provide rules for the relative position of words within sentences.

QUEST uses the following logical functions :-

(i) The affirmation. Example: A
This simple declaration of a keyword A means: search for all documents in which the word A appears.

(ii) The negation. Example: .NOT.A
This means: search for those documents in which A does not appear.

(iii) The conjunction. Example: A.AND.B
This means: search for those documents in which the word A and also the word B appear.

(iv) The inclusive disjunction. Example: A.OR.B
This means: search for those documents in which A or B appears, or both A and B appear. The expression '.OR.A' is thus equivalent to 'A'.

(v) The exclusive disjunction. Example: A.XOR.B
This means: search for those documents in which A or B appears, but not both. The exclusive disjunction is not an essential part of the language since it can be formed from a combination of .OR., .NOT., and .AND. but it may on occasion be preferable to use it directly.

These logical functions can be combined in more complicated relationships by employing the usual rules for hierarchical orders of precedence. Brackets may be used in the normal way to assist in defining the function. The following is an example of a more complicated expression:

((A.OR.B).NOT.C).AND.(D.AND.E).AND.((F.OR.G).XOR.H)

Positional relationships are used to prescribe the degree of proximity of one word with another within a sentence. Thus (A.AND.B(5).AND.C(1)) means that B must be five words or less after A and C immediately following B in the same sentence. Positive and negative numbers can be used: (A.AND.B(+5,-3)) means that B must be five words or less after A or three words or less before A. Positional relationships apply only within the same sentence so that in order to specify that words be in the same sentence it is only necessary to use some large number. Thus (A.AND.B(± 999)) means in effect look for all documents in which A and B are in the same sentence. Since it is often desirable to look for strings of words in succession a recent version of the QUEST language enables this to be done by simply writing down the words in sequence : so that A B C is equivalent to A.AND.B(+1).AND.C(+1).

As a practical example of the use of QUEST consider a search for the phrase 'justice of the peace'. All documents containing this phrase could be found by use of one of the following statements:

(a) JUSTICE.AND.PEACE
(b) JUSTICE.AND.PEACE (+999,-999)
(c) JUSTICE.AND.PEACE (+3)
(d) JUSTICE.AND.OF(+1).AND.THE(+1).AND.PEACE(+1)
(e) JUSTICE OF THE PEACE

Statement (a) retrieves those documents in which the words 'justice' and 'peace' occur, whereas statement (b) retrieves only those documents in which both words occur in the same sentence. For statement (c) to be true the documents have to contain the word 'justice' followed within three places by the word 'peace'; it would therefore recover a document in which the phrase 'justice and peace' occurs. Statements (d) and (e) are alternative ways of searching for the exact phrase 'justice of the peace' : only documents containing these four words in the given order would be retrieved.

6. The Search Process : Typical Examples

In order to illustrate the range of searching techniques available by computer it is helpful to discuss some typical examples.

One of the simplest searches is to look for those documents which contain a string of words in sequence such as : Atomic Energy Authority. The presence of only one of these words in a document is quickly found by consulting the concordance: it is convenient to use the search program to look for a number of words in sequence. Two possible ways of expressing the search enquiry are :

(a) ATOMIC.AND.ENERGY.AND.AUTHORITY
(b) ATOMIC ENERGY AUTHORITY

The first statement finds those documents in which the three designated words appear whatever their order or position, whereas the second statement retrieves only those documents in which the words occur in succession as shown. The result of the search is to give titles of those documents which satisfy the queries. Twenty-three sections or schedules of Acts of Parliament satisfy search (a) whereas seventeen satisfy search (b).

RESULTS OF COMPUTER SEARCH

SEARCH NUMBER 30

THE SEARCH QUERY IS SATISFIED BY 6 DOCUMENTS -

ATOMIC ENERGY ACT 1946 CH 80 SEC 5.

ATOMIC ENERGY ACT 1946 CH 80 SEC 12.

ATOMIC ENERGY AUTHORITY ACT 1954 CH 32 SEC 2.

ATOMIC ENERGY AUTHORITY ACT 1954 CH 32 SEC 3.

SCIENCE AND TECHNOLOGY ACT 1965 CH 4 SEC 4.

SCIENCE AND TECHNOLOGY ACT 1965 CH 4 SCHEDULE 3.

Fig.8

///

SCIENCE AND TECHNOLOGY ACT 1965 CH 4 SEC 4. J473

EXTENSION OF RESEARCH FUNCTIONS OF ATOMIC ENERGY AUTHORITY. M473

(1) THE FUNCTIONS OF THE UNITED KINGDOM ATOMIC ENERGY L474
AUTHORITY SHALL INCLUDE THE UNDERTAKING OF SCIENTIFIC RESEARCH L475
IN SUCH MATTERS NOT CONNECTED WITH ATOMIC ENERGY AS MAY, AFTER L476
CONSULTATION WITH THE AUTHORITY, BE REQUIRED BY THE MINISTER OF L477
TECHNOLOGY, AND SECTION 2(2) OF THE *ATOMIC ENERGY AUTHORITY + L478
ACT 1954* SHALL APPLY AS IF ANY SUCH RESEARCH WERE RESEARCH INTO L479
MATTERS CONNECTED WITH ATOMIC ENERGY. L480
(2) THERE SHALL BE DEFRAYED OUT OF MONEYS PROVIDED BY PARLIA+ L481
MENT ANY INCREASE ATTRIBUTABLE TO SUBSECTION (1) ABOVE IN THE L482
SUMS PAYABLE UNDER SECTION 4(1) OF THE *ATOMIC ENERGY AUTHORITY + L483
ACT 1954* OUT OF MONEYS SO PROVIDED. L484
(3) SECTION 3(6) AND (7) ABOVE SHALL HAVE EFFECT IN RELATION TO L485
ANY ACTIVITIES CARRIED ON OR TO BE CARRIED ON BY THE UNITED L486
KINGDOM ATOMIC ENERGY AUTHORITY BY VIRTUE OF THIS SECTION AS L487
IF THE AUTHORITY WERE A GOVERNMENT DEPARTMENT. L488
///

Fig.9

Thus there are six documents which contain the words 'atomic', 'energy' and 'authority' without them constituting the phrase 'atomic energy authority'. For example, section eleven of the Atomic Energy Act 1946 is concerned with restrictions on the disclosure of information relating to atomic energy without the general authority of the Minister.

A more elaborate legal enquiry might be concerned with retrieving those documents which set out the powers of the Atomic Energy Authority, particularly its power to undertake research and development. A suitably framed search enquiry written in the QUEST language is the following :-

```
(ATOMIC.AND.ENERGY.AND.AUTHORITY).AND.(POWER.OR.POWERS.
OR.DUTY.OR.DUTIES.OR.FUNCTION.OR.FUNCTIONS).AND.
(RESEARCH.OR.DEVELOPMENT)

.PRINT.12
```

Here the computer is asked to search for conjunctions of three bracketed groups of words. The first consists of a conjunction of 'atomic' and 'energy' and 'authority' ; the second of a disjunction of the words 'power', 'duty' and 'function' and their plurals; and the third a disjunction of the words 'research' and 'development'. These keywords have been chosen to express in QUEST language the concepts in the legal enquiry. The statement .PRINT.12 instructs the computer to print the text of the documents which satisfy the enquiry as long as there are no more than twelve of them.

The computer print-out obtained as a result of this search is shown in Fig.8. Six documents satisfy the search enquiry and the corresponding J cards, or titles of the documents, are listed. The Atomic Energy Authority Act 1954 ss.2-3 and the Science and Technology Act 1965 s.4 are directly relevant to the enquiry for they deal with the powers and functions of the Authority, the powers and duties of the Minister in relation to the Authority and the extension of the research and development functions of the Authority. The other three documents are not relevant and may be classed as 'false drops', that is to say the context of the documents is of little or no use in answering the query. This proportion of relevant to irrelevant material is typical of the results obtained in computer searching. Since there were no more than twelve cited documents in this search the text was printed out and the listing of the shortest one, Science and Technology Act 1965 s.4, is shown in Fig.9.

A third type of search is exemplified by the search for a definition of a word or phrase. The functions of the Atomic Energy Authority, for example, are extended by the text of Fig.9 to "include the undertaking of scientific research in such matters not connected with atomic energy as may .....". Suppose it is necessary to search the Acts to ascertain whether the phrase 'scientific research' is there defined. In computer terms this means looking for the word 'scientific' followed by the word 'research' and if this phrase is found, searching further to establish whether the context of the phrase is that of a definition. In Acts of Parliament words are commonly defined by saying that they 'mean' something (a restrictive definition), or 'include' something (an extensive definition) or are to be 'construed' in a certain way, or for the purposes of the Act are 'deemed' to be something else. A suitable enquiry framed in the QUEST language might therefore be as follows :-

///

SCIENCE AND TECHNOLOGY ACT 1965 CH 4 SEC 6. J525

SUPPLEMENTARY. M525

(1) IN THIS ACT =SCIENTIFIC RESEARCH= MEANS RESEARCH AND L526
DEVELOPMENT IN ANY OF THE SCIENCES (INCLUDING THE SOCIAL SCIENCES) L527
OR IN TECHNOLOGY. L528
(2) NOTHING IN THIS ACT SHALL PREJUDICE OR AFFECT ANY POWER L529
TO AMEND OR REVOKE THE CHARTERS OF ANY RESEARCH COUNCIL, L530
OR ANY POWER OF HER MAJESTY TO GRANT NEW CHARTERS, OR AFFECT THE L531
OPERATION OF ANY AMENDMENT MADE OR CHARTER GRANTED AFTER THE L532
PASSING OF THIS ACT. L533
(3) THE ENACTMENTS MENTIONED IN SCHEDULE 4 TO THIS ACT L534
ARE HEREBY REPEALED TO THE EXTENT SPECIFIED IN THE THIRD COLUMN L535
OF THAT SCHEDULE, WITH EFFECT IN EACH CASE FROM SUCH DAY AS HER L536
MAJESTY MAY BY ORDER IN COUNCIL APPOINT. L537
///

Fig.10

```
SCIENTIFIC RESEARCH.AND.(MEAN.OR.MEANS.OR.MEANING.OR.
INCLUDE.OR.INCLUDES.OR.DEFINED.OR.DEFINITIONS.OR.
DEEMED.OR.CONSTRUED)(+12,-12)
```

This statement means that the phrase 'scientific research' has to be found in conjunction with one or more of a group of words - the disjunction inside the brackets - and these words have to be no more than twelve words away from the word 'research'. In practice it has been found that this search procedure will successfully retrieve the definitions in the atomic energy Acts.

The above expression is however rather clumsy so a new function, the .DEFINE. instruction, has been introduced into the QUEST language specifically for searching for definitions. In the jargon of computer programming this is a 'macro' instruction and in terms of this macro the search enquiry can be written simply :

```
.DEFINE.SCIENTIFIC RESEARCH
```

and the computer will accept this as an instruction to carry out the search for a definition described above. The .DEFINE. instruction is the beginning of a more natural language for searching statutes.

The result of this particular search enquiry is retrieval of just one document, the Science and Technology Act 1965 s.6, and the text of this section as printed by the computer is shown in Fig.10. It can be seen that for the purposes of the Act 'scientific research' means research and development in any of the sciences (including the social sciences) or in technology.

7. Discussion and Conclusion

The preliminary version of the STATUS programs is now working satisfactorily and carrying out the tasks expected of it with reasonable efficiency. Nevertheless, there are many minor, and perhaps some major, improvements that should be made : for example, the present programs do not include any efficient method of up-dating the dictionary and concordance when the text is amended. The programs are still in the research and development phase and not yet in a fit condition to carry out a regular production job on a large volume of text.

Two further developments in particular are required: the programs need to be tested on a larger data base; and the QUEST language needs to be extended. The atomic energy legislation, though very suitable for a feasibility study, is too small to provide an adequate test of program efficiency or of searching methods. The next stage requires a text of about one or two million words which would allow the programs to be developed to the point where they could deal with the twenty million words or so of all Acts of Parliament currently in force.

In developing the QUEST language, the aim should be to make it as close as possible to natural language so that it can be used by those unfamiliar with computers. There are a number of simple improvements that could readily be provided: for example, the use of word roots and the automatic addition of plurals and synonyms to key words. But beyond that there is a need to construct a search language which is similar to the language naturally used by lawyers, and to develop the search process in such a way that it possesses the flexibility and richness of a conventional search by an experienced searcher.

There is much yet to be done but it is possible to draw a number of conclusions from the results of the STATUS project so far. Firstly, they point to the value of computer-generated aids to the text. The dictionary is by itself of help since it indicates what words are present in the text and with what frequency they appear; and the index provides immediate access to the context of each keyword. But it is the concordance which is probably of most value : many simple searches can be carried out manually using this alone. Secondly, the results of computer searching have proved encouraging. Any string of words can be located quickly and unerringly, and complicated logical and positional relationships between words can be examined. These searches often retrieve irrelevant material; but experience shows that if the search enquiry is prepared with care, then all the relevant material can be obtained. This is not to say that computer searching is easy or simple, for it certainly is not. It must be a fundamentally difficult task to express in suitable keywords the substance of a legal concept, if for no other reason than that different draftsmen use different words and possess differing styles. But the experience gained so far indicates that computer searching, if not by itself the complete answer to textual searching, is at the least certain to be a most valuable aid to the conventional process, providing the searcher with precise, untiring, and rapid assistance with his work.

Perhaps the most valuable feature of computer analysis of Acts of Parliament will be the insight it provides into the process of legal communication. No doubt for completely effective searching the computer programs and interrogation language have to be further developed to better match the way in which Acts are drafted: but they also show ways in which the Acts can themselves be improved so that the computer can search them in a more effective manner. As in so many other areas of computer application, perhaps the most valuable contribution the computer can make is to provide a completely new look at the legislative communication process. It is fitting to quote Professor Horty on this point. In discussing computer searching of the Pennsylvanian code he wrote[4] :

> "Once we .... let in the mathematicians, logicians, computer scientists and others to take a look at what the lawyer does, legal research, the legal profession and the whole process of the administration is never going to be the same again."

The preliminary results of the STATUS project amply confirm this point of view.

8. Acknowledgements

The authors wish to express their grateful thanks to the many persons, too numerous to list here, who have assisted the progress of this work with their advice, encouragement and interest. Special acknowledgement is due to: Dr. K.V. Roberts for his valuable help with the early programs; Dr. K.W. Browning who provided much practical programming assistance; Dr. M.D. Poole who was responsible for the on-line version of the search program and has given much helpful advice; Miss N.E. Couchman who prepared the Key-Word-In-Context program; and last, but by no means least, Miss R. Ford who helped in innumerable ways with the development of the programs.

9. References

1. G.B.F. NIBLETT, "The Computerization of the Statute Book", The Computer Bulletin, 12, June (1968).

2. W.S. RHYNE, "Law Research by Computer. Pamphlet No.4", World Peace Through Law Centre, Geneva, Switzerland, (August 1966).

3. A.S. FRAENKEL, "Legal Information Retrieval", Advances in Computers, 9, 113 (1968).

4. J.F. HORTY, "Keywords in Combination Approach to Computer Research in Law with Comments on Costs", Modern Uses of Logic in Law, 54 (March 1962).

5. W.B. KEHL, J.F. HORTY, C.R.T. BACON and D.S. MITCHELL, "An Information Retrieval Language for Legal Studies", Communications of the ACM, 4, 380 (1961).

6. C.F. TAPPER, "British Experience in Legal Information Retrieval", Modern Uses of Logic in Law, 127 (December 1964).

7. C.F. TAPPER, "Use of Computers for Lawyers", J.Soc.Public Teachers of Law", 8, 261 (December 1965).

8. See K.S. POPE, "The Lawyer and the Computer", Seminar paper presented at the Fifteenth Legal Convention of the Law Council of Australia, Brisbane, Australia (July 1966) at p.16.

9. See H. STREET and F.R. FRAME, Law Relating to Nuclear Energy, Butterworths (1966) at page v.

10. H.P. LUHN, Amer.Documentation, 11, 288 (1960).

Available from

HER MAJESTY'S STATIONERY OFFICE

49 High Holborn, London, W.C.1
13a Castle Street, Edinburgh 2
109 St. Mary Street, Cardiff CF1 1JW
Brazennose Street, Manchester M60 8AS
50 Fairfax Street, Bristol BS1 3DE
258 Broad Street, Birmingham 1
7-11 Linenhall Street, Belfast BT2 8AY

or through any bookseller.