Lib sc. 6; 1969; PAPER C.

Vital Role of Depth Classification in a System for Document-Finding: A Trend Report.
(Non-conventional methods in document retrieval. 8).

G Bhattacharyya, *Documentation Research and Training Centre Bangalore* 3.

[S .rveys some of the significant works of the period 1945–1968 on the .se of a sche. ie for classification in designing a mechanised Syste.i for Doc i ient-Finding (= SDF). Concludes that a harmonious co ibination of the potentiality of the New Methodology of designing a scheme for depth classification developed in DRTC with the ability of a co ry .ter for fast search, in the design of a SDF wo ld be able to meet more efficiently the document needs of specialists.]

0 Introduction

01 "SPEEDS"

According to the Five Laws of Library Science (23), the ultimate objective of a System for Document-Finding (= SDF) is to disclose to its users all that it has in store, in response to his specifications. This service should essentially be pin-pointed, exhaustive, and expeditious. That is why, the ideal system as suggested by the Five Laws, has been described by the acronym SPEEDS (= System for Pinpointed, Exhaustive, and Expeditious Dissemination of Subjects) (1).

02 SUBJECT APPROACH AND SDF

Of all the specifications brought forward by a user, it is the subject specification that is responsible for the complex nature of a SDF. Again, of all the points-of-approach, it is the subject approach that is paramount. That is why, the history of the design of SDF is mostly the history of the organisation of the subjects embodied in documents.

In response to a subject-specification by a user, the responsibility of a SDF is to disclose to him all the references to

1 Documents coextensively conforming to that specification;
2 Documents embodying subjects of greater extension than

the specification of the user, but having a substantial portion devoted to it;

3 Documents embodying subjects of smaller extension than the specification given by the user but dealing with some aspect or part of it; and

4 Documents embodying collateral subjects which contain some information, though indirectly, on the subject specified by the user (25).

03 BEGINNING OF SDF

To discharge this responsibility, a SDF started its journey from Panizzi's "Index of Matter" (19). An entry for the Index of Matters was to be made under a word occurring in a title and indicative of the subject treated in it. However, in the course of its evolution, a SDF ultimately attained the form of a Classified Catalogue based on an analytico-synthetic scheme for classification and supplemented by a subject index constructed by following the rules of Chain Procedure (20).

04 MECHANISATION OF SDF

The introduction of the mechanisation of a SDF was first thought of to gain speed in its operation. Mechanisation, in its initial stage, was not employed to the latest evolved form of SDF, but to the primitive form — that is, "Index of Matter" — which by that time changed its name into "Keyword Index" based on the title. Non-structured "keywords" soon proved inadequate. As an alternative came Thesaurus — a list of structured compulsory keywords. The need for structuring originated from the necessity to build up groups. In this way, a classification element got introduced, and simultaneously the work of assig ing a subject heading to a document became a process of analysis.

05 CLASSIFICATION AND MECHANISATION

Of late, attention has turned to exploit the advantages of a scheme for classification in a mechanised SDF. Much work has already been done in this area. Results of experiments have shown that to exploit the capabilities of a computer to the fullest benefit

1 The theoretical basis of the scheme for classification should be thoroughly sound; and

2 It should be able to analyse the subject to a great depth.

This paper attempts to survey the works on the use of a scheme for classification — that is, a scheme of classes fitted with notation (24), in designing a mechanised SDF.

Ranganathan's work on this subject is not included in this survey. However, the essence of his ideas may be summarised

in his own words as follows: "Machinery or no machinery for retrieval, classification is a necessity — not just a superficial classification but depth classification." (21)

06 BUSH'S PROPHECY

A survey as proposed in Sec 05 should take note of Vannevar Bush's prophecy in 1945 about the future SDF (3). He called it "Memex". It may be described as follows:

A memex is a device in which an individual stores his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to the memory of its user. It consists of a desk. On the top are slanting transluscent screens, on which material can be projected for convenient reading. At one end is the document stored. The matter of bulk is well taken care of by improved microfilm. Most of the memex contents are purchased on microfilm ready for insertion. And there is provision for direct entry, dry photography being employed. There is, of course, provision for consultation of the record by the usual scheme of indexing. If the user wishes to consult a certain book, he taps its code on the keyboard, and the title page of the book promptly appears before him, projected on to one of his viewing positions. Frequently-used codes are mnemonic, so that he seldom consults his code book; but when he does, a single tap of a key projects it for his use. All this is conventional, except for the projection forward of present-day mechanism and gadgetry. It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of a memex.

It may be noted that the essential feature of the memex is an associative index; and finding of documents from it is through codes. These two factors together are suggestive of the introduction of a scheme for classification in designing a SDF.

1 Speculation

11 FIRST VIEWS

Views about the feasibility of developing a mechanised document finding system based on a scheme for classification was expressed first, probably, in 1948 (28). The Royal Society's Working Party on Mechanical Indexing remarked that UDC was a system "ready for machine coding". In making this remark, the Working Party primarily took into consideration the detailed schedules of UDC.

12 BRISCH'S REPORT

E G Brisch reported an experiment with UDC using punched

card equipment, such as sorter, collator, and tabulator (2). The experiment was based on a fixed field card design. This design allotted 51 columns for all the facets of a Class Number. The rest of the columns were allotted to other items of information. The length of a main UDC number was not allowed to exceed 5 digits. Similar restrictions were imposed on common auxiliary facet members. On the basis of this experiment, it was concluded that

1 The restriction on the length of Class Number would prove unhelpful in organising entries for a collection of any significant size in a specialised field; and

2 The use of the UDC notational system in its normal form was practically impossible with sorter and collator, because the ordinal value of the digits in a UDC class number was quite different from what these equipments were ordinarily equipped to sense.

13 WORKING PARTY'S CONJECTURE

The Royal Society's Working Party noted these difficulties in handling UDC on punched card equipment. At that time, electronic data processing machines were in their first stage of development. The Working Group speculated that in future the electronic machines might prove suitable for the purpose (28).

2 Decision of Mechanising an Abstracting Service

In 1949, the authority of the *Meteorological and geoastrophysical abstracts* (= MGA) decided to use the UDC for arrangement, coding, and ultimate mechanisation of the storage, selection and indexing of the world's meteorological literature (27). The actual use of machines, however, was deliberately delayed for ten years for the following reasons:

1 Fixed field card design, then in vogue, could not accommodate the variable length of UDC numbers, names of authors, titles of serials and other bibliographical elements, without using special codes for each;

2 Existing tabulating equipment was not capable of providing the quality and accuracy of presentation and arrangement desirable for a publication which would be used in libraries for years; and

3 Available experience was insufficient to provide the intellectual input for designing a system.

3 Experiment with Edge-Punched Card

In 1951, W R Rushton reported an experiment with UDC using edge-punched cards (29). On the basis of this experiment,

it was concluded that to derive advantages, there was need of modifiying the Class Numbers, because the "codes triangle" cards could not cope with more than nine-figure numbers, nor with symbols other than numerals.

31 Compromise

311 *Experiment on "Power-Samas" Sorter*

In 1958, R H Richens reported a mechanised system using UDC that was under trial at the Commonwealth Bureau of Plant Breeding and Genetics (26). He noted the advantages of a scheme for classification in devising a mechanised document retrieval system. He said "For them [Scientists] a ready-made classification, even though unsatisfactory on several scores, is likely to be of great use". For his system, the 65-column "Power Samas" cards of ten different colours were used. The cards were hand-punched; and a "Power Samas" sorter having a selective sorting attachment was used for selection. The difficulties described by Brisch were experienced while devising the system. Therefore, a compromise was sought in the card-design. In the layout of the card each of the main subjects in which a specialist in plant breeding and genetics is usually interested, was assigned a separate panel of five columns. The subject "Agriculture" was made an exception to it. For "Agriculture" three five-column panels were used. Another five-column panel was allocated to "General". While punching, the first two digits of these main UDC numbers were instructed to be omitted. For a common auxiliary of form, another five-column panel was allocated; it was entitled "Tags". The remainder of the card was occupied by a three-column panel for representing the decimal number of the country to which a paper refers, a three-column panel for representing its year of publication, a two-column panel for indicating the number of the volume of *Plant breeding abstracts*, in which a summary of the article had appeared, and a final one-column panel entitled "Class" in which a hole was punched in position 1 to indicate that the card is a subject card.

Thus, the difficulties mentioned by Brisch were avoided by

1 Restricting the length of class number;
2 Omitting the first two digits of a main UDC number; and
3 Not using the auxiliaries except for that denoted by "form".

32 Early Stages of Development of MGT

In the early stages of the mechanisation of the *Meteorological and geoastrophysical titles* (= MGT) several compromises were sought to facilitate sorting of UDC numbers on punched cards by

1 Not using any special digit except parentheses; and

2 Filling out to three digits by adding zeros to the right of each inter-decimal point group (5).

4 Until 1960

Until 1960, the advancement in the field of equipment was relatively slow. Attempts at handling UDC were largely made on sorters, collators, and tabulators. Therefore, breaking through the limitations of punched cards was not feasible without sacrificing a considerable length of the UDC number. Vickery said, "Terms are codes on to cards for mechanisal search. Some systems have used UDC class numbers directly for this purpose, but in general, it is an extravagant misuse of coding space. Machine-searched document card systems typically cover only a restricted subject field and two of the factors leading to long UDC notations are felt as grave disadvantages: the repetition of the 'main class' in the class number and the longer numbers of 'common facets'. It seems likely, therefore, that UDC notation is not useful for mechanised search systems although the schedules themselves provide a valuable source of terms and hierarchies for such system." (32).

5 Gremas

"Gremas" is an acronym for Generic Retrieval by Magnetic Tape Storage. Robert Fugmann's report (12) shows that the preliminary work on the development of the "Gremas" system dates back to 1955. "Gremas" is an analytico-synthetic scheme for classification used in the documentation of organic chemistry. In 1958, the analysis of documents and the storing on magnetic tape was begun. In 1960, the store was released for literature searches. On an average, 8–10 subject specialists were continuously engaged in literature analysis, making searches and developing the system. By 1964, 150,000 structural formulae of organic chemical compounds and of groups of compounds had been stored as well as several thousand organic chemical reactions. The machines IBM 705 and IBM 1401 were used. Thirty-two different machine programs were in operation for storage and retrieval.

On the basis of his experience, Fugmann concluded, "A perfect solution of all problems connected with the storing and retrieving of information in Chemistry can be expected only from a consistently analytico-synthetic classification, the fundamentals of which are laid down in Ranganathan's *Prolegomena to library classification* . . . .

> "In documentation the adequate field of application of modern computers have not always been found till now. Frequently, computers have exerted fascination merely by their high working speed. Therefore, they have frequently been applied to the mere speeding up of traditional

working steps without previously attempting to improve the conceptual basis of documentation. Then the only success to be observed was that the unsatisfactory result of the machine selection was available — not as formerly after several hours — but already after several minutes.

"On the other hand, many attempts have been made to use computers for tasks which would have been better left to human intelligence. The main task of documentation is to answer the generic questions of a user, ie, to retrieve references, which are pertinent to a group of related concepts. Relationship among concepts means conformity in their essentials. However, what is essential can only be stated by expert judgement. Be the mere processing of the data as precise and quick as it may in a documentation, defects which are caused by omitting definition and classification of the concepts cannot be compensated by any machine work. This mental activity always has to precede the mere processing of the data by machine.

"Therefore, in the era of computers, classification must not play a minor part. On the contrary, the theory and practice of classification must be promoted strongly in the future if we wish to achieve real progress in the documentation field."

6 Leslie's Effort

Another example of the use of a classification system in automatic retrieval was reported by W H P Leslie in 1961 (17). He described how technical abstracts were coded, filed and retrieved in the particular circumstances at the National Engineering Laboratory, East Kilbridge, where a DEUCE high-speed computer was available. The author was not satisfied with the UDC for indexing abstracts. Therefore, he evolved a similar system designed to fit the instrumentation field. The report pertained to the very early stage of development of the work. At that time there were only two or three other laboratories in Great Britian where some kind of elementary exercises on this subject were done.

7 Automation for Classification

Since 1962, an automatic classification method for biology has been in operation (31). Reporting on this method, Martin Scheele observed that automation as well as classification were facing the great task of managing knowledge. Only by close cooperation of the old discipline of classification with the modern systems of automation of the problems of "knowledge arrangement for research, education, teaching, and documentation could be solved in a satisfactory manner", he said. "Modern automation", according to him, "unlocks entirely new possibilities for classification." He further remarked that it became imperative to "think-through" the bases of classification anew; and it was for that reason that classification research had become so important today. He introduced his description of the method by saying

"We started from the consideration that not all documentation problems can be tackled in one working process. Today, frequently difficult tasks are started instead of first really seeking to solve simpler problems with the aid of automation. In documentation too, one should not take the second step before the first (put the cart before the horse!), since this way valuable time is lost and unnecessary difficulties and expenses are incurred. The first step in documentation is the evaluation (classification and processing) of titles and all the other easily processed bibliographic data about the publications."

By 1964, over 50,000 "terms" were classified and cards were punched to create a thesaurus for classifying automatically. The classification used was a faceted system, distinguishing three main divisions: Chemical compounds, Organisms, and Properties. The latter division had 40 individual facets.

Scheele noted the advantages of classifying automatically as follows:

"In our method each occurring word has to be classified by humans but once. If there is any additional occurrence of the word, in the titles of the publications, then the computer classified automatically, with the aid of its thesaurus. This results in two decisive advantages.

"First, we can employ a heavily faceted classification, where frequently several more-than-one-digit notations can be allotted to the individual words. This would be too complicated if humans and not a computer would classify.

"Secondly, the entire title material can be reclassified automatically if one changes the notations of the magnetic-tape-thesaurus. In this manner also concordances between various classification systems can be achieved, while it would be practically impossible to classify anew several hundred thousand titles through a human work team."

8 Mechanised Control of Meteorological Literature Using UDC

In 1964, M Rigby reported on the experiments in mechanised control of meteorological and geoastrophysical documents using UDC (27). He mentioned that after ten years of in-house experience with the use of UDC, subject headings, serial titles, monthly, annual and cumulative indexes, and the concurrent development of more sophisticated machines, electronic computers, and above all the availability of programmers with some knowledge of documentation problems, the time was ripe to begin a step-by-step developmental programme aimed at complete mechanisation of indexing. Some of the mechanised operations which had proven possible with the use of punched card or magnetic tape input, had been as follows:

1 Beginning in October 1960, a list of 20,000 subject headings with equivalent UDC numbers used in volumes 1 to 10 of the *Meteorological abstracts and bibliography* was punched on IBM 026 Print-punch machine. The cards were arranged alphabetically for one printout, and sequentially by UDC numbers for a systematic or thesaurus-type printout. Copies were made by using IBM 407 tabulator. Since 1962, this list was supplemented by another 10,000 headings for volumes 11 to 13.

2 During 1962 and 1963, the UNIDEK indexing system (a systematically arranged presentation of entries according to UDC numbers) was used in *Meteorological and geoastrological titles*. It was programmed to be run on an IBM 407 Tabulator or an IBM 1401 Computer. The programmes were designed by Robert R Freeman.

3 A multi-access computer system was developed during 1962 and 1963 with the help of programmers of CEIR, Inc., Arlington, Virginia, for the purpose of compilation, indexing, and printout of selective bibliographies on any desired subject with inclusion of complete abstracts, if desired.

4 Starting from 1962, attempts were made to determine costs, and show the feasibility of mechanising the UDC schedules for the purpose of editing, updating, printing, studying, comparing, translating, indexing etc.

5 During the months January to May, 1964, Robert R Freeman developed a prototype mechanised system of selective retrieval, indexing and dissemination of current titles, using for demonstration a sample of 360 titles in the field of Oceanography selected from 2,200 titles in all fileds of geophysics and astrophysics which were received by *Meteorological and geoastrophysical abstracts* during the last quarter of 1963. The system was based on UDC.

Among the projects needing further exploration, Rigby mentioned about the automatic assignment of UDC numbers from keywords in title and abstracts.

On the basis of the experience with all the experiments mentioned above, Rigby concluded that there were no limitations to the direct use of UDC in computer retrieval or printout systems, which were not inherent in any other system.

91 Restriction in the Notational Plane

Reporting his experience of the mechanisation of the *Meteorological and geoastrophysical titles*, Robert R Freeman (5) said, "where some advantage accrues to humans from the use of a particular notation, it should be used and the burden of manipulating it correctly should be on the machine and the programmer". He further said that in the case of computers with their large memories and stored programme control, the problem was greatly reduced in magnitude. Fixed fields were unnecessary; and notation which yielded class numbers of the length common to UDC presented no problem, he said. He was of the opinion that while it admittedly required some additional machine time, it was even possible to compensate for the difference between recommended filing order and machine collating sequence by programming. However, even in 1964, there were certain restrictions imposed on the notational plane of UDC.

For, parenthesis and the slash were still the only other symbols employed. Class numbers related by the colon were treated as though they were separate unrelated numbers assigned to a document. He planned to use fifty-one printing characters of IBM 1401 to lift-up these restrictions. In programming he would instruct the machine to look for certain characters in the input data, substitute other characters for them, sort the data, and then resubstitute the original characters or substitute still others before printing the output. On the use of notation, he said that there must be a mechanism which could control the vocabulary of search, enable the search to be broadened or narrowed, and suggest alternate path. The notation of UDC along with a well-conceived thorough A/Z index was uniquely well qualified to meet these requirements, he said. He suggested the need for a fresh look at the use of data processing equipment in conjunction with classification systems in the light of advancing technology. "Classificatory notation" he said, "appears to be a valuable tool for use in machine information systems when used in conjunction with other tool."

92 Emphasis on Classification for Automatic Documentation

921 Meyer-Uhlenried's Views

Speaking of the advantages of classification over keywords, K H Meyer-Uhlenried observed in 1964 (18):

> "Taking into account the large number of the different keywords in use, the pure coordinate indexing method, that is, the unstructured purely alphabetic order of equivalent terms, is no longer optimal, because there are too many wrong combinations if the indexing is to have the necessary depth and because the degree of the order, which can be reached in this way, is very low ··· This dilemma in automatic documentation led some of the systems to a real crisis. As an alternative, the vocabularies used have been modified at first and compulsory lists of keywords, so-called Thesauri, have been compiled. These thesauri were more or less structured. Giving this structure to the originally completely open word lists shows, on the one hand, the necessity to build up groups — that is, a clear classification element — and on the other hand, an essential problem of automatic documentation is shifted to the analysis."

922 Gardin's Views

In 1964, J C Gardin used the term 'facet' to denote the groupings of a "special" classification in a particular field of work (13). With reference to this denotation he designed schemes for classification into two genera: "Faceted" classification and "Free" classification. In a so-called "faceted" classification, the groupings are based on functional analogies with reference to specific observational contexts (eg, Agents of a reaction, Products of extraction etc). In the so-called "Free" classification,

classes mark essential affinities of a more general validity, ie, context-free (eg, chemical bodies, materials, etc). While speaking of the advantages of these two genera of classification in relation to their use in computers, he observed as follows:

> "An important characteristic of faceted systems, from this point of view, is that the data can be presented according to a fixed model, corresponding to the *a priori* division of the field in a given number of zones — the facets — in which the indexing terms are necessarily arranged in accordance with the analysis of the documents. This order can be used with advantage in the internal organisation of the data in the memory so as to reduce the time of search and calculation ··· The possibility of recording the information in a "fixed field" is in this case, an important detail of the automatic exploitation of faceted systems. The characteristics of this method of organisation can moreover be formulated in terms of a very simple model which in some ways constitutes the archetype of all faceted classifications, independent of their field or their particular contents. From the notion of the model we pass easily to that of a *general programme* for the recording and automatic searching of data expressed in accordance with the norms of the model, a programme which is applicable to the exploitation of all faceted documentary systems, whatever the field may be."

He reported that a programme of the above nature had been undertaken by Mme M Renaud at the Section for Automatic Documentation in collaboration with Mme Saville at the Bull Machines C. A first experiment had been done on the Gamma 30 to determine the time and approximate cost of the application.

His observations on the use of "Free" classification in computers is also worth noting. He said that the systems which had been called "Free", were not adequate to the same methods of automatic treatment. "The functional equivalent of division into facets, for the expression of logical structures," according to him, "is then a more or less complex system of syntactic indicators connected to the index terms ("roles", "interfixes", "links", etc) whose manipulation in the computer imposes different internal organisations and exploratory routines."

923 HUTCHIN'S VIEWS

Discussing the need for the introduction of classification into an index in the context of automatic document selection without indexing, W J Hutchins said (15) that it is necessitated by the restrictions placed on the number of descriptors permitted for each document.

> "Without restrictions the index file would rival the document collection itself in size. Thus, each descriptor designates a large class of denotata communicated in the document; it stands in a generic relationship to them. Another descriptor in the index may refer to some of these denotata but not to others (class intersection) and a third may refer to all and also to some others (class inclusion). Thus, descriptors can be

ranked in hierarchies according to the domain of their reference. If relevant documents are retrieved by a search request formulated in one descriptor, others may be retrieved by descriptors related to it in such a hierarchy. If no classification were provided, these other relevant documents would be missed."

93 Caless and Kirk's Efforts

In September 1967, T W Caless and D B Kirk (4) reported an experiment on an application of UDC to machine searching. The experiment was conducted on a collection of seismological documents. The system was a flexible one and could be adjusted to any desired level of exhaustivity of indexing. A technique of manipulating UDC numbers was adopted for the purpose of this experiment. An IBM 1401 computer was used. On the basis of their experience, the authors concluded that the use of UDC as a language for machine retrieval was quite feasible.

94 Computer Manipulation of Notation

In September 1967, Theodore C Hines wrote on the computer manipulation of classification notation (14). He furnished a practical filing rubric with explanation of machine filing concepts. Remarking on the suitability of the notations of the Library of Congress and Dewey Decimal Classification for computer search, Hines said:

> "Since both systems would require schedule look-up for machine search, and since this involves, for the computer, comparisons rather than straight manipulations, neither system has any particular machine advantage over the other, at least in the absence of further data than has yet appeared. Each may have some advantage for particular purposes. LC may take more sorting time, but this is conjectural in the absence of large-scale tests."

About UDC he maintained that it might certainly offer more machine searching possibilities.

95 Evaluation of UDC as Indexing Language for A MSDF

The National Science Foundation sponsored a research project for the evaluation of the UDC as the indexing language for mechanised reference retrieval system (= MRRS) in 1965. Mrs Pauline Atherton was appointed the General Project Manager and Robert R Freeman the Project Supervisor.

951 REPORT 1

Freeman in his Report 1 on the project (8) defined the aims of this project as follows:

1 To develop a complete English language version of UDC in both hierarchical and alphabetical arrangement in machine-readable form;

2 To develop techniques for automatic file maintenance and photocomposition of UDC editions;
3 To develop a computer-based reference retrieval system which would use UDC as its indexing language;
4 To collect a set of UDC-indexed document files in machine-readable form in various subject areas; and
5 To conduct tests with the aid of experimental user groups which will lead to an evaluation of the UDC in the desired context.

The earlier work of Malcolm Rigby (*See* Sec 8) became largely responsible for conceiving this project. Rigby's work with UDC presaged virtually every aspect of this project.

952 REPORT 2

In his Report 2 (9) submitted in February 1966, Freeman gave an account of the progress made till then in the project. They had succeeded in
1 Collecting the schedules of UDC in different languages;
2 Translating some of the schedules;
3 Maintaining a UDC file on IBM 1401;
4 Experimenting with automatic alphabetic indexing of UDC schedules;
5 Selecting equipment and formulating rules for key-boarding the UDC into machine readable form; and
6 Taking initial steps towards collection of UDC-indexed documents and developing a retrieval system for test and evaluation purposes.

In relation to the last mentioned field of work, he reported that investigations had centred around two subject areas: Metallurgy and Nuclear Science.

953 REPORT 3

In his Report 3 (7) submitted in October 1966, Freeman reviewed the problem of managing a classification such as UDC as an example of the broader class of problems known in the system analysis and data processing field as "file management". The uses of data processing equipment for the creation, maintenance, manipulation, and display of files were discussed.

954 REPORT 4

Report 4 relating to the Project (30) was submitted in April 1967. In this report, Martin Russel and Robert Freeman included a case study of the adoption of UNIDEK by *Geosciences abstracts* — a computer-compiled systematic subject index based on UDC. While evaluating UDC in this context, they observed thus:

1 With all its drawbacks UDC was the best indexing language widely known or used for the earth sciences; and

2 The greatest weakness in computer-aided indexing by UNIDEK was the lack of a mechanically searchable file of all possible UDC numbers from which headings could be selected by a matching process.

They concluded: "The fact that several foreign language journals both primary and abstract, include UDC numbers, raises the interesting possibility of mechanically producing English-language indexes of journals published in "difficult" languages from the UDC numbers alone".

955 REPORT 5

Report 5 (11) relating to the Project was submitted in September 1967. In this report, Robert Freeman and Pauline Atherton have furnished a model of contemporary mechanised retrieval systems, and the characteristics of indexing languages used therein. In this context they have developed a rational basis for the use of UDC. The use of UDC as an indexing language with the Combined File Search System has been demonstrated. On the basis of their experience they observed, "For mechanised retrieval systems, the advantage would appear to rest with the indexing language with a notation which not only uniquely identified the descriptors, but reveals relationships among them." They concluded: There is no longer any doubt that the UDC can be used as the indexing language in a mechanised system ··· The UDC as it presently exists, probably cannot function as efficiently in a mechanised system as an indexing language designed specifically for machine processing, but no barriers exist to the successful use of the UDC in either a batch processing or interactive mode." They have mentioned that under certain conditions, it may be desirable to use UDC. Some of these conditions are the existence of large files already indexed by UDC, staff already trained for its use, and extensive international use, or exchange of materials of the system. Finally, they have discussed how UDC might be used as a query-language in a typical retrieval system of the near future in which the user interacts directly with the computer-stored document reference file.

956 REPORT 6

Report 6 relating to the Project (6) was submitted in April 1968. This report embodies an account of an experiment concerned with the processing of a set of 25 questions against a computer-stored file of 9,159 document references in the field of ferrous metallurgy. These references represent the 1965 coverage of the Iron and Steel Institute (London) information

service. The American Society of Metals (= ASM) had processed earlier many questions against a data-base which contained many of the same documents. The experience of the ASM provided the basis for evaluation of system performance characteristics and analysis of system failures. The results of this experiment, have led to the following conclusions and recommendations:

1 The use of UDC as the indexing language in computer-based retrieval system may be encouraged: for, the performance characteristics derived for the system are entirely satisfactory.

2 The search strategy in a computer-based retrieval system using UDC needs to take into consideration actual distributions and combinatorial probabilities of the UDC numbers used for indexing. Use of such feedback from the system should be a considerable aid to the system operator in attempting to optimise the balance of levels of recall and precision according to his chosen policy.

3 According to the above policy, no search would request a UDC class number to be present which had not been used as to index at least one document, even though the class may exist in theory in the UD schedule. Thus, if a UDC class number is found not to have been used in the system, it is up to the search analyst to specify whether or not the number for a more general concept should be used.

4 To effect improved performance of a mechanised system using UDC it is recommended that

41 The use of a general class to describe a few subjects collectively be avoided, and each subject be indexed separately;

42 The policy of indexing only the several most important subjects be avoided, and each distinctly identifiable topic which appears in the title and abstract of a document be indexed;

43 The use of an indexing work sheet which reminds the indexer to consider each of the possible facets, which may be anticipated, of a subject be introduced.

5 To meet the requirements and capabilities of computer-based systems continued revision of UDC according to the principles and techniques of faceted classification is necessary.

957 REPORT 7

Report 7 relating to the Project (10) was also submitted in April 1968. The report describes an experimental system for remote direct access to files of computer-stored information which has been indexed by UDC. The system has been named AUDACIOUS, an acronym for *Au*tomatic *D*irect *A*ccess to *I*nformation with the *On*-Line *UDC S*ystem. The data-base for the experient consisted of references from a single issue of *Nuclear*

Science Abstracts. The *Special Subject Edition of UDC for Nuclear Science and Technology* was also stored in the computer so that users could discover how to translate their questions from natural language to logical statement containing UDC numbers.

AUDACIOUS is the first on-line interactive retrieval system in which one of the widely used traditional schemes for classification, UDC, has been used. The results of this experiment have led to the following conclusions and recommendations:

1 While UDC was the tool in the case of AUDACIOUS, the success of the experiment may be generalisable to other tools, such as the Dewey Decimal and Library of Congress Classifications.

2 In future, investigation should be made to explore the use of UDC in conjunction with a suitably detailed thesaurus. UDC might be used to rapidly narrow the portion of the file to be searched to a small size, the thesaurus then being used for detailed interaction with that subject of the file.

3 It needs to be investigated whether users prefer a natural language or a numeric or other code for interacting.

4 The users' viewpoints should be considered carefully in designing an interactive retrieval system. A system which is a technical success can fail to impress a user in many areas.

The authors, Robert Freeman and Pauline Atherton have conjectured that the improvements during the past few years in communications and computer technologies strongly indicate that networks of libraries and information centres, whose resources are linked electronically, will be feasible in the not-too-distant future. They have said that the users will be able to conduct searches by means of a dialogue with the system, with access to distant as well as geographically nearby files of information. Such networks need not be confined within national borders. As to the manner of indexing that would serve adequately for users who do not share a common natural language, they have submitted two solutions:

1 Use of the language in which largest volume of literature is written — that is, English; and

2 Use of a form of indexing that is not dependent on natural language, such as UDC.

96 Conclusion

961 Nature of a Query to be Answered

The nature of a query that a SDF is now expected to satisfy has been illustrated in the example cited by G H Hutton (16). He writes:

> "Design is largely a series of choices. Each selection of a material or component is made on the available criteria from the known field of

products. The evolution of criteria and comparison and selection are fundamental design activities ··· Suppose, however, that a designer required to know which manufacturers supplied a metal window of the following description:

TYPE:	double hung vertical sliding sash
SIZE:	5 ft 8 in × 4 ft 0 in wide
MATERIAL:	aluminium alloy HE9
FINISH:	Stain anodised to grade M35 to BS 1615: 1961 with strippable lacquer protection
OPERATION:	gear-operated with plastic-coated opening handle
GLAZING:	suitable for sealed double glazing units

"This description is neither exhaustive nor unusual in its requirements and yet ··· no classification in common use and no catalogue distribution service could readily supply the names (if any) of manufacturers producing such a window.

"This example illustrates the difficulty facing designers in selecting a product to satisfy known requirements. The information which is readily available is often so general as to be almost useless, while the designer's greatest need — for specific information — is frequently unsatisfied because of the time and effort necessary to find this information."

962 KIND OF SCHEME FOR CLASSIFICATION NEEDED

The use of the depth version of Colon Classification, designed according to the New Methodology (22), can analyse the subjects to the degree desired by the design engineers.

963 CAPABILITIES OF A COMPUTER

A computer can be made to perform the function of matching a query-specification with a subject-specification.

"All the inputs once stored in the computer can, through proper programming, be combined with each other at will and can again be written in any desired sequence. The best known advantage of the modern methods are their large storage capacity and their operational speed, in addition to the error-free reproducibility of all data and values that once have been stored correctly. Together with their multidimensionality and variability these are the most important computer properties, exceeding human capacities by far." (30).

964 COMBINATION OF CLASSIFICATION AND COMPUTER

Consideration of the above three factors leads one to conclude that a harmonious combination of the last two — that is, the potentiality of the New Methodology of subject analysis, and the capabilities of a computer — in designing a SDF would be able to meet the requirements of the design engineers to the fullest extent. It is necessary to look at the problem afresh and try to explore this possibility as early as possible.

965 DRTC COMPUTER PROJECT

The results of preliminary experiments in this regard done in DRTC have been found satisfactory. A noteworthy

feature of this experiment is the synthesis of Class Numbers by the computer. Reports on this experiment are being published in a series of articles.

97 Bibliographical References

1 Sec 01 BHATTACHARYYA (G). Selective dissemination of information (SDI): Comparison with SPEEDS. (Paper submitted for discussion at the 34th FID Conference and International Congress on Scientific Information (Moscow) (1968)).

2 Sec 12 BRISCH (E G). Adaptation of the UDC form of notation to punched card technique. (*In* Royal Society Scientific Information Conference. Report. 1948. P 690–3).

3 Sec 06 BUSH (Vannevar). As we may think. (Atlantic M. 176; 1945; 101–8).

4 Sec 93 CALESS (T W) and KIRK (D B). Application of UDC to machine searching. (J doc. 23; 1967; 208–15).

5 Sec 91 FREEMAN (Robert R). Computers and classification systems. (J doc. 20; 1964; 137–43).

6 Sec 956 ——. Evaluation of the retrieval of metallurgical document references using the Universal Decimal Classification in a computer-based system. 1 April 1968. (American Institute of Physics, UDC project, report. AIP/UDC-6).

7 Sec 953 ——. Modern approaches to the management of classification. 1 October 1966. (American Institute of Physics, UDC Project, report. AIP/UDC-3).

8 Sec 951 ——. Research project for the evaluation of the UDC as the indexing language for a mechanised reference retrieval system: An introduction. 1 October 1965. (American Institute of Physics, Documentation Research Project, report. AIP/DRP UDC–i).

9 Sec 952 ——. Research project for the evaluation of the UDC as the indexing language for a mechanised reference retrieval system: Progress report for the period July 1, 1965 — January 31, 1966. 1 February 1966. (American Institute of Physics, Documentation Research Project, report. AID/DRP UCD-2).

10 Sec 957 —— and ATHERTON (Pauline). AUDACIOUS — an experiment with an on line, interactive reference retrieval system using the Universal Decimal Classification as the index language in the field of nuclear science. 25 April 1968. (American Institute of Physics, UDC project, report. AIP/UDC-7)

11 Sec 955 —— and ——. File organisation and search strategy using the Universal Decimal Classification in mechanised reference retrieval systems. 15 September 1967. (American Institute of Physics, UDC Project, report. AIP/UDC-5).

12 Sec 5 FUGMANN (Robert). Experiences with a faceted classification in organic chemistry using computers. (Proceedings, International Study Conference on Classification Research. 2; 1964; Elsinore; 341–62).

13 Sec 922 GARDIN (J C). Free classifications and faceted classifications: Their exploitation with computers. (Proceedings. International Study Conference in Classification Research. 2; 1964; Elsinore; 161–8).

14 Sec 94 HINES (Theodore C). Computer manipulation of classification notation. (J Doc. 23; 1967; 216–23).

15 Sec 923 HUTCHINS (W J). Automatic document selection without indexing. (J Doc. 23; 1967; 273–90; Sec 1·5).

16 Sec 961 HUTTON (G H). Product analysis by co-ordinate index. (Aslib Proc. 20; 1968; 171–2).
17 Sec 6 LESLIE (W H P). Automatic retrieval of technical information. (Aslib Proc. 13; 1961; 145–53).
18 Sec 921 MEYER-UHLENRIED (K H). Classification and automatic documentation. (EURATOM research). Proceedings, International Study Conference on Classification Research. 2; 1964; Elsinore; 331).
19 Sec 03 PANIZZI (A). Minutes of evidence. (*In* Olding (R K), *Ed.* Readings in library cataloguing. 1966. P 28).
20 Sec 03 RANGANATHAN (S R). Classified catalogue code. Ed 5. 1964.
21 Sec 05 ——. Depth classification, tools for retrieval, and organisation for research. (*In* Ranganathan (S R), *Ed.* Documentation and its facets. 1963. Chap P2).
22 Sec 962 ——. Design of depth classification: Methodology. (Lib sc. 1; 1964; Paper A).
23 Sec 01 ——. Five Laws of library science. Ed 2. 1957.
24 Sec 05 ——. Prolegomena to library classification. Ed 2. 1957. Sec 1283.
25 Sec 02 ——. Theory of library catalogue. 1938. P 79–80.
26 Sec 311 RICHENS (R H). An abstracting and information service for plant breeding and genetics. (*In* Casey (R S) and others, *Ed.* Punched cards. Ed 2. 1958. Chap 17.)
27 Sec 2 8 RIGBY (M). Experiments in mechanised control of meteorological and geoastrological literature and the UDC schedules in these fields. (Rev Int Doc. 31; 1964; 103–6).
28 Sec 11 13 ROYAL SOCIETY SCIENTIFIC INFORMATION CONFERENCE. Report. 1948. P 160.
29 Sec 3 RUSHTON (W R). Application combinee des cartes perforees et de la classification decimale universelle pour la bibliographie. (Rev Int Doc. 18; 1951; 31–40).
30 Sec 954 RUSSELL (Martin) and FREEMAN (Robert R). Computer aided indexing of a scientific abstracts journal by the UDC with UNIDEK: A case study. 1 April 1967. (American Institute of Physics, UDC Project, report. AIP/UDC-4).
31 Sec 7 SCHEELE (Martin). Significance of automation for classification. (Proceedings. International Study Conference on Classification Research. 2; 1964; 368–77).
32 Sec 4 VICKERY (B C). UDC and technical information indexing. (Unesco Bul Lib. 15; 1961; 126–38).