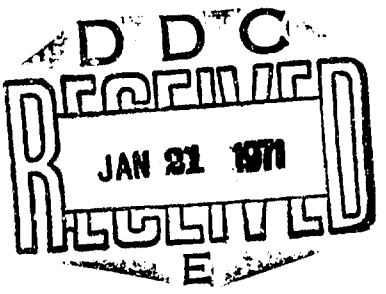
STRATEGIES FOR MANIPULATING UNIVERSAL DECIMAL CLASSIFICATION RELATIONSHIPS FOR COMPUTER RETRIEVAL

Final Report
December 1970



Prepared for:

U.S. Air Force Office of Scientific Research
Washington, D.C.

BIOLOGICAL SCIENCES COMMUNICATION PROJECT
THE GEORGE WASHINGTON UNIVERSITY MEDICAL CENTER
2001 S STREET, N.W., WASHINGTON, D.C. 20009
Telephone (202) 462-5828

FINAL REPORT

U.S. Air Force Office of Scientific Research
Contract #F 44620-68-C-0035

STRATEGIES FOR MANIPULATING UNIVERSAL DECIMAL CLASSIFICATION RELATIONSHIPS FOR COMPUTER RETRIEVAL

by

T.W.Caless, A.C. Foskett,
D. Langridge, J. Mills and
J.M. Perreault

December 1970

Biological Sciences Communication Project
The George Washington University Medical Center
Department of Medical and Public Affairs

Charles W. Shilling, M.D., Director
Biological Sciences Communication Project

CONTENTS

PREFACE

Although this report is to serve as the Final Report to Contract F 44620-68-C-0035, it is only "final" in that the research funds were exhausted. The thorough development of any General scheme (or even a Special scheme) would normally require a considerably larger technical staff for a longer period than was possible on this project. Additionally, exhaustive testing is required of all technical developments, preferably with "live" document collections. This in itself can be quite arduous and time consuming when dealing with the non-rigorous natural language.

Ideally, the systematic organization of any information collection requires extensive use of the services of both subject specialists and contributors to the development of library classification. Because of this, provisions were made during contract negotiations to engage the part-time services of Jean M. Perreault, Jack Mills, Derek Langridge and A.C. Foskett as Consultants to the project. Each of these gentlemen supplied the project with material when called upon, and when it appears in this report, it has been properly credited.

ABSTRACT

Summarizes research findings of Principal Investigator and part-time Consultants who conducted an investigation of the Universal Decimal Classification (UDC) as an indexing language for computer retrieval. Problem areas were identified (notation, schedules, application), new developments reviewed (combining Precedence devices, subject analysis matrices, guidelines for schedule revision, pilot schedules, parenthesis-free notation for searching), and a broad overview and history of the UDC are presented in ten papers. Faceted classification techniques applied to the UDC are recommended throughout the investigation for better consistency in application.

INTRODUCTION

Throughout this investigation, the Principal Investigator was keenly aware of the disruption that new developments to the Universal Decimal Classification (UDC) would cause to existing library collections organized with this system. Yet the research had to proceed, not with the intent to disrupt, but with the intent to critically examine, to identify deficiencies, and to resolve the problems. Otherwise, there is little hope that the UDC could prosper in the future. It must be remembered that the opponents of the UDC frequently voice *valid* criticisms and it is their objections that we cannot overlook. Of out-dated and non-professionally prepared schedules, we can be critical—of the fundamental concepts and potential of the UDC, we can be intrigued. From a practical aspect, what other scheme exists that could serve as the international bibliographic standard?

It was to this ultimate end that the research was directed. First, to resolve the uncertainties in the notation and its application; next, to establish technical guidelines for schedule revision; and finally, to develop pilot schedules which could be field tested and later formally submitted to the International Federation for Documentation (FID) for approval and authorization.

Simultaneously with the above, a careful study of computer search strategies was underway. It was apparent early in the research that unless one resolved the uncertainties in the fundamentals, there could be no progress in the development of the search strategies. It appeared that a logical technique which would linearize a complicated indexing string composed of UDC notation would be an excellent way to effectively search the file. It was at this point in the research that the funds were interrupted.

HISTORICAL REVIEW

THE UNIVERSAL DECIMAL CLASSIFICATION (A.C. Foskett)

In 1894 two Belgians, Paul Otlet and Henri Lafontaine, conceived of a universal index to recorded knowledge, to which people all over the world might contribute, and which would in its turn be available to all. It was of course not feasible to think of alphabetical arrangement of index terms in such a situation, since it was envisaged that contributions would be received from nations with many different languages; a systematic arrangement using a notation was essential to the success of the enterprise. Consideration of the possibilities showed that a scheme existed which might serve the purpose: Dewey's Decimal Classification, then in its fifth edition. Otlet and Lafontaine sought and obtained Dewey's permission to modify his scheme in order to make it suitable for bibliographical rather than shelf arrangement. This was done largely by the superimposition of several devices for showing bibliographical form, the language, the date, and so on. They then classified some thousands of articles in time for the First International Conference on Bibliography, held at their suggestion in 1895. As a result of this conference, an international organization, the IIB (Institute International de la Bibliographie) was set up to manage the index. This body later became the International Federation for Documentation (FID).

With the development of the index, it proved necessary to develop also the classification scheme, and the latter was published in 1905 with the title, "Manuel du repertoire universel bibliographique,"

emphasizing its function as the means of arrangement of this one index. However, once published the scheme was adopted by many libraries and government organizations in Western Europe, and its use spread gradually throughout the world, despite the fact that it was available only in French.

The Great War of 1914-1918 was a heavy blow for the index, and in the early 1920's it was abandoned. By this time the classification scheme was sufficiently well established in its own right to warrant a second edition, published in French over the years 1927-1933. Otlet and Lafontaine supervised the revision of the Humanities and Social sciences; Frits Donker Duyvis, seconded the Natural sciences while at the Dutch Patent Office. A third edition in German was begun in 1934; publication of this was interrupted by the Second World War, and was not completed until 1952, a separate index being published in three volumes 1951-1953. The fourth edition, in English, began publication in 1943 but is not yet complete; it is hoped that the completed schedules will be available by the end of 1969, but those published in 1943 will not have been revised. As these include Class 5 (Science) and Class 0 (Generalities), there will obviously remain a substantial amount of work to do before it can legitimately be claimed that a complete up-to-date edition exists in English. Other editions in progress include the Japanese and Spanish and a new German edition, but at present the latest full edition completely available is the now somewhat out-of-date first German edition.

A classification scheme which is not available to potential users is unlikely to succeed. The Universal Decimal Classification (UDC) has therefore been published in a number of languages in abridged editions; the English abridgement, originally based on the usage of the Science Museum Library in London, whose Director, S.C. Bradford, was an ardent supporter of UDC, is now in its third edition, dated 1961. This is the edition most widely used in English-speaking libraries. In recent years 'medium' editions have also been prepared; the first of these, in German, was published in 1967. The distinction between Full, Medium and Abridged editions appears to be a pragmatic one; Mills (see pages 6-12) has been unable to find any theoretical basis for the degree of abridgement, which varies from subject to subject. Overall, there is intended to be an approximate ratio of 10:1 between Full and Abridged, with the Medium falling at the geometric mean, but this is not necessarily found in any particular subject.

The revision procedure for UDC is a source of both strength and of weakness. Drafts of new schedules are usually prepared by users working in libraries in the subject area, and thus reflect the current needs of users of the literature. They are submitted to international subject committees, consisting of librarians and other interested parties from all over the world. Once approved by these committees, drafts go to the Central Classification Committee, which is the body having direct responsibility within the International Federation for Documentation for the UDC. When approved by the CCC, drafts are published as P-Notes, which are open to comment by any UDC user for a period of four months. It is evident that to be accepted, a draft has to undergo scrutiny by a number of well-qualified experts, and is thus likely to satisfy the needs of the revision area subject-wise, though one must enter a "caveat" here that many of the experts are not expert in classification theory, and new schedules may not stand up well to a searching examination on this ground. On the other hand, the process of seeing a draft through the various committees involved can be tedious, since the work has normally to be done through correspondence, while objections to a P-Note can lead to even lengthier delays. Two years would be good progress from beginning work to final acceptance; the schedule for aerospace science took eight years before it gained final approval.

P-Notes to which no objections are raised become part of the official schedules and are published in the next issue of Extensions and Corrections to the UDC; this is published at six-monthly intervals and cumulates progressively. All extensions and corrections affecting the first German edition, up to the end of 1964, are incorporated in one set of volumes; a further volume covers 1965, 1966, and 1967, while it is necessary to consult the latest issue for changes since the end of 1967.

To discover the latest state of a particular schedule, it is necessary to consult not only the latest full edition published (which may be the German) but also the series of Extensions and Corrections; in addition, it is as well to study recent P-Notes to see whether any further amendments are presently under discussion. Lack of funds often means that proposals are not translated from their original language, normally English, French or German. The combination of having to look in a number of places for the

latest schedule and having to be able to translate freely from French and German may prove to be something of a deterrent to the potential user.

The scheme itself bears, in outline at least, a considerable resemblance to its parent, the Dewey Decimal Classification (DC), though there has been one major change: the transfer of Philology from its place in DC between Social sciences and Natural sciences to a more satisfactory position within a wider group Language and Literature. This resemblance means that in many ways the overall order in UDC no longer reflects modern thought: for example, Psychology is found as a subdivision of Philosophy; Transport engineering is separated from Railway, Highway engineering by Hydraulic engineering and Public health engineering; Nuclear, atomic and molecular physics follow all the rest of Physics instead of preceding it. Two studies of UDC commissioned by UNESCO and published in 1961 were highly critical of the scheme, largely on the basis of its poor overall order. Major changes have been proposed, such as the development of a new group of 'bridge' subjects, e.g., Communication, to fit into the place between Social sciences and Natural sciences left vacant by the transfer of Philology. Little progress on these developments is evident, and it seems that they have aroused considerable opposition among the more conservative users of UDC. As with any information retrieval system, there is a conflict between past practice and present needs, to which there is not an easy solution.

Once the detailed structure of the UDC is studied, however, the resemblance to DC ceases, largely through the use in UDC of a number of notational devices. The purpose of the notation in a classification scheme is often misunderstood; it is thought that the notation *is* the classification, whereas it is in fact merely the means whereby the arrangement dictated by the classification can conveniently be seen and used. The notation must be capable of reflecting the systematic arrangement completely, and this leads to certain problems. UDC, like other classification schemes, is intended to be used in a precoordinate manner: that is, composite subjects, consisting of several concepts in association, are treated as units, rather than split up into their elements as is the practice with postcoordinate indexing systems. It is, of course, impossible to foresee all the composite subjects which may arise, and modern classification schemes are therefore synthetic: that is, they list the individual concepts which are found within a given subject area, and give the individual classifier the means whereby he can associate these concepts to reflect the composite subjects found in documents. It is at this point that notational problems can arise; while it is a simple matter to devise a notation which will reflect an enumerated order, it is much more difficult to devise one which will reflect not only the enumerated order but also permit the insertion of composite subjects, not enumerated but synthesized, in their correct places in the overall arrangement. For example, it must be possible to specify not merely Nuclear Reactors, but also Fast Reactors (i.e., a type), Reactor Coolants (a part), Reactor Fuel Elements (a part), Reactor Fuel Element Cans (part of a part), Deterioration of Reactor Fuel Element Cans (a process occurring within a part of a part), Reactor Control (an operation on a reactor), Fast Reactor Control, Fast Reactor Fuel Elements, Deterioration of Fast Reactor Fuel Elements—and any other combination of concepts which may occur. To do this without restriction requires that each concept should have its own piece of notation which may be combined unambiguously with any other.

As indicated above, it is found that the concepts occurring within a particular subject area can be grouped into sets—parts, processes, operations, types, etc.—which are called facets. A concept within the context of a facet is known as a focus; if the analysis of the subject has been completed satisfactorily, it will be found that foci within the same facet are mutually exclusive: that is, they cannot be combined. For example, we may have Fast Reactors and Thermal Reactors; Fast Reactor Fuel Elements and Thermal Reactor Fuel Elements; but we cannot have fast thermal reactors—the combination is meaningless. The notation must therefore permit the unambiguous combination of foci from different facets, in such a way that the result is a meaningful translation, into symbols having an ordinal value, of the original composite subject.

UDC notation contains a number of devices which are intended to permit synthesis. Many of these relate to 'common' facets—those like Time and Place, Language and Bibliographical Form—which may apply to any subject or document. There are certain symbols which are used for the more specific kinds of

synthesis to be found within the context of particular subjects: these are the colon, hyphen and point 0 (: , - , . 0). Of these, the hyphen and point 0 are used to introduce the facets within a subject, and their meaning as facet indicators depends on their context. For example, in Philology -56 means Syntax, in Engineering -56 means Torque Control; in Physics .083 means Methods of Measurement, in Literature .083 means Editing. By contrast, the colon is used to join pieces of notation from anywhere in the complete schedules; it has the significance 'in relation to' and nothing more. We may find it used to divide a subject directly (genus-species), e.g.,

635.965	Indoor Plants (horticulture)
635.965:582.675	Indoor Anemones

or it may be used to combine foci from different facets, e.g.,

635.965:632.38	Virus Diseases of Indoor Plants

or it may be used to indicate a more diffuse relationship such as the essentially *ad hoc* phase relationships of influence, comparison, bias or exposition, e.g.,

635.965:697.38	Effect of Hot-Air Central Heating on Indoor Plants
635.965:747	Indoor Plants for Interior Decoration

It must be remembered that the synthetic devices in UDC were developed some time before modern classification theory had distinguished between the several different kinds of relationship. Recent research has shown that their lack of precision is a severe handicap in a computer-based retrieval system. In addition, the use of the hyphen and point 0 is by no means standardized throughout the scheme, so that users who wish to develop new schedules may be in some doubt as to the best way to allocate notation.

In addition to the common facets, which are found in every classification scheme, and the special signs of relationship just described, UDC contains three other kinds of notational symbols. The first of these are the signs of aggregation, the slash and the plus (/ and +). In a classification scheme, division should always be proximate, i.e., it should not omit any steps between a genus and its species subdivisions. Unfortunately, Dewey in his original outline did in fact omit many such steps, which were therefore not represented by notation. For example, in Zoology we find that the schedule goes direct from the general heading to the specific headings Invertebrata and Vertebrata, omitting the intermediate heading Animal Taxonomy. The use of the slash permits us to insert such omitted headings, e.g.,

59	Zoology
591	General Zoology (i.e., processes facet)
592/599	Systematic Zoology. Animal Taxonomy
592	Invertebrata
596	Vertebrata
	etc.

In effect, the slash shows consecutive extension, to give notation which includes all the topics which are found within the aggregate heading originally omitted.

There are also cases of aggregation where consecutive extension is not involved. There are two ways of treating Science and Technology, for example: the first is to group all sciences together and all technologies together; the second is to group together each science and its dependent technology. Both of these are legitimate groupings, but it is not possible to use both in the linear context of a classification scheme. The use of the + makes it possible to denote those aggregates which are not catered for by the grouping in the scheme itself. Dewey chose the first of the two alternatives above, i.e., all sciences are

grouped, as are all technologies. UDC therefore follows the same pattern, but can group specific sciences with their technologies by use of the +, e.g., 539.1+621.039 Nuclear science and technology. The problems of using the + to denote such aggregates have been discussed by Foskett (see pages 21-23).

The second of the special signs of relationship is the .00, used to introduce the so-called Point of View numbers. These represent a kind of common facet in which concepts as economics, operation and management are to be found. The organization of this facet, or perhaps rather a set of facets, is often in conflict with facets enumerated within subjects in the main schedules, and the title 'point of view' is a misnomer; only in one case can it be claimed that this is the correct designation, and that is the relatively recent addition of .00 colon to introduce the author's point of view. For example, a work on dialectics written from the Marxist point of view could be given the notation 162.6.00:335.5, in which dialectics is represented by 162.6, Marxism by 335.5.

The third symbol used with a special meaning is the apostrophe. This is mainly used at present in subjects such as Chemistry and Metallurgy, where the idea of subject synthesis as opposed to notational synthesis is found. For example, in Chemistry, Sodium is 546.33, Chlorine 546.13; the Chemistry of Salt is thus 546.33'13. It has been argued that to restrict the apostrophe to this use is to waste one of the limited number of notational symbols available (limited by the fact that they must appear on a standard typewriter keyboard). However, if other schedules are developed along these lines, it should prove possible to retain the apostrophe for this purpose.

One other symbol has been used in UDC notation: square brackets. An important feature of a synthetic classification scheme is the citation order, i.e., the order in which the elements of a composite subject are to be cited in the notation. For example, Place is normally subsidiary to all the facets of a particular subject, and the notation for Place therefore appears towards the end of the complete notational symbol. There will, however, be topics where Place is more significant than this, for example in Political parties or even Politics generally; in such cases we need to be able to manipulate the notation in such a way as to bring the notation for Place nearer the beginning of the total symbol. The notation for Place does in fact permit this, as

539.1+[621.039:65.011.56]
[539.1+621.039]:65.011.56

This use of square brackets combined with a standard order of application of devices is also discussed by Perreault (see pages 18-21).

The need for revision is of course recognized, and to avoid the confusion which might arise if notational symbols were reused with changed meaning a policy of 'starvation' is in effect. This means that a new schedule uses different notation from the old, and that the old notation is left unused for a minimum of ten years, by which time it is assumed that it may if necessary be reused without too much confusion. For example, until 1961 the schedule for Particle Accelerators used the notational base 621.384.61 and 621.384.62; a new schedule was developed using 621.384.63 to 621.384.66 and leaving the previous notation vacant. By 1972 it will be possible to reuse the previous notation if it should prove desirable. The method can lead to a lengthening of the base notation in some cases, but in general this will not be a serious problem, particularly as the notation for detailed composite subjects will in any case be long.

It will be clear from this brief account that UDC as it stands at present suffers from a number of defects, which may be tabulated:

1) The overall order of subjects reflects in many ways an age now past. Many new subjects, and even some not so new, have had to be incorporated in unsatisfactory locations in relation to their parent disciplines.
2) The schedules reflect a mixture of enumeration and synthesis which in many cases leads to ambiguity and unsatisfactory order.
3) The notation is often not sufficiently precise to stand up to the demands of a tightly organized retrieval system.

4) The revision procedure tends to be slow and erratic. Schedules are revised only when there is an expressed demand and someone willing to do the work, so that the overall impression is very uneven; some parts are up to date, others have clearly not been revised fully for many years.
5) There is no clear guidance as to how new schedules should be compiled when the need arises. Many recent schedules have been developed along analytico-synthetic lines, by users aware of recent classification theory; others appear to be compiled along the old, haphazard, lines of enumerative classification. There appears to be little consistency in the way that the notational devices are used. The document published by the FID dealing with revision procedures refers only to the administrative problems.
6) The publication program, particularly of the full editions, leaves much to be desired.

To set against these negative features, there are a number of positive aspects which suggest that UDC can continue to play an important role in future information systems:

1) It is possible to revise most of the unsatisfactory features within the present framework; indeed, many recent revisions, for example the schedules for Nuclear science, Nuclear technology, and Seismology, are completely in line with modern theory and would need little amendment in any overall revision.
2) The publication procedure can be transformed by computer techniques, so that it would be possible for any user to have at all times a completely up-dated set of schedules.
3) The notation, once it is applied consistently and precisely, will lend itself well to computer operation.
4) UDC is used by a great many libraries around the world, and will almost certainly continue to be used very widely. It contains a great deal of invaluable enumeration, the work of a considerable number of experts, which is unequalled by any other scheme.
5) It has the backing of a substantial international agency, the FID.
6) It is used by a number of abstracting and indexing services either as the basis of their arrangement, or else as an additional aid to users.
7) It is now the official classification scheme in scientific and technical libraries in the USSR. A number of revision proposals have come from Soviet Union in recent years, for example the use of .00: to represent author's point of view, and a detailed schedule for Economic geology, this latter being still at the P-Note stage.

It would seem that perhaps the main problem that has faced UDC in the past has been lack of funds to support adequate reorganization. Only in the very recent past has the FID been able to employ even a small permanent Secretariat, and most of the revision work has been done by librarians in their spare time. The British Standards Institution (BSI), the body responsible for publishing the English edition, was unable to employ more than one Technical Officer for UDC until 1968, when a grant from England's Office of Science and Technology Information (OSTI) enabled BSI to treble their staff immediately. It is perhaps an indication of the possibilities that many of the most successful recent revisions, for example that in Nuclear sciences and technology, have come from large-scale organization where it has been considered part of the official work of the library staff to develop the necessary schedules. Outside this kind of situation, UDC has a curiously amateur air; Frits Donker Duyvis who as General Secretary of FID was probably the greatest single influence on UDC for a period of some thirty years, was never paid by the FID. (He was employed by the Dutch Patent Office.) A definite, concerted effort by all interested parties might well yield substantial improvements in the scheme as a whole.

FUNDAMENTALS

FULL, MEDIUM AND ABRIDGED EDITIONS OF UDC (J. Mills)

These descriptions refer to the amount of detail enumerated in the classification schedules. The significance of detail in a classification is that precision in searching an index depends on the precision with

which the subjects of documents can be defined. This precision in description does not, however, depend only on enumerated detail. The degree of synthesis allowed is also very important. The interaction of these two features is considered in detail later.

1. UDC editions and their degree of detail

1.1 The basic official edition is the Brussels expansion of 1927-33 (the 2nd international (full) edition, in French), supplemented by the subsequent German, English and other full editions (although only the German is as yet complete).

1.2 The first *abridged* edition appeared in 1922, in German; the first English abridged edition appeared in 1938. The significant Trilingual Edition (1958), together with the various individual national abridgements established a rough standard as to detail, about 1:10 in relation to the full editions. Although a main motive in producing abridged editions has been the desire to produce a simple introduction to the system and one less daunting to beginners than the full edition, it seems likely that the delays in producing the full editions have also been a strong reason, i.e., the abridgements have been seen as a stop-gap pending the appearance of the full schedules.

1.3 The *Medium* editions which have begun to appear in the last decade aim at a one-volume edition, substantially more detailed than the Abridgements (of the order of 3:1) but still considerably less detailed than the full editions. Again, it seems likely that the continued delays in producing complete full editions has been an important reason for their appearance.

1.4 Although the various reductions appear to approximate to two fairly constant proportions (1:3 and 1:10) it does seem that there are no precise rules governing the production of UDC schedules of varying size, either as to the total size of a given edition or to the balance between classes in one edition. So some exploration of the possibilities would seem to be desirable if explicit objectives are to be met and consistency observed.

2. Specificity in UDC

2.1 Historically, UDC developed from the Dewey DC as a response to the need for maximum detail in specification. It sought to achieve this by developing not only detailed enumeration of subclasses with generic hierarchies but also *synthesis* as a major method of signifying compound classes. This was an alternative to the elaborate enumeration of such classes (i.e., classes combining terms from two or more facets—e.g., *Insect pests in storage of grain crops*) which is severely limited by the impossibility of pointing out *all*, or even many of the possible combinations which can be formed between elementary concepts (i.e., concepts reflecting one facet only).

2.2 Synthesis implies, inevitably, the prior analysis of subjects into relatively elementary classes (Insects, Pests, Storage, Grain. . .). Such conceptual analysis leads, of course, to facet analysis, although the full implications of this were not apparent to the early compilers of UDC and are still not as apparent as they should be to some participants in the revision and expansion of UDC.

2.3 Alongside the need to develop conceptual analysis was the need to develop a notation able to implement the synthesis on which achievement of really detailed subject-specification depends. Again, this led inevitably to faceted notation and the various auxiliary devices and the colon demonstrate this.

3. Specificity in index languages

3.1 The ability to be precise in index descriptions is essentially the resolving power of the index language in class definition. A broad classification cannot focus on as small a class as a detailed one.

3.2 Of the two major parameters by which index performance is measured–Recall and Precision–the latter is entirely dependent on the ability of the index language to provide high resolving power. Recall depends on indexing policy (as to exhaustivity,–i.e. the thoroughness with which a document's information content is described) and searching policy (the more broadly a search is made the greater the chance of retrieval). A failure or inadequacy in exhaustivity of indexing can be compensated for by broader searching. But nothing can compensate for inadequacy in the specificity of indexing–without specific index descriptions. Precision is limited absolutely.

3.3 It should be noted that this argument rests on an interpretation of specificity as being the *generic hierarchical level* at which a concept is described by the index language. For example, a document dealing with Insect pest attack on Potatoes in clamp storage may be analysed as a compound formed by the intersection of terms from 3 different facets (in the class Agriculture)–a process (disease) a crop, and an operation (i.e., an action performed *on* a crop). Exhaustive indexing of this topic must recognize all 3 facets; but this could be done at different hierarchical levels; e.g.,

Crops	Biological Processes	Agricultural Operations
Field crops	Disease	Storage
Root Crops	Caused by Pests	Clamp Storage
Potatoes	Insect Pests	

If this were indexed as Potatoes–Insect pests–Storage, the description would be exhaustive but not specific. Specificity implies recognition of all the facets (categories) of concepts and each one at its most specific hierarchical level (in the generic hierarchy making up the facet). In the example above, this is represented by taking the bottom term from each of the three hierarchies. A criticism of this interpretation of specificity is considered later on. (Note also that any reduction in exhaustivity automatically effects a reduction in specificity (but not vice-versa, as we have seen). For example, if this subject were indexed as clamp storage of Potatoes, the absence of any recognition of the Processes facet is akin to indexing "All processes considered"–we say that the facet is 'diffuse'.

3.4 It should be noted that specificity is a relative quality in a practical retrieval situation. For example, in a special collection on Transport administration, a question on *Passenger services* would be a broad one. Greater specificity in such a collection would entail classes such as *Passenger booking facilities at airports*. But in a general collection, a question on Passenger services would be relatively precise and a lower level of specificity in index description might achieve an equally high precision in searching. In performance measuring, this situation is referred to as the 'generality ratio' and must be accounted for in assessing precision.

3.5 It follows from the above that resolving power (the ability to describe a subject specifically) is dependent on two different facilities in an index language–the specificity in its individual facets (really, a question of how detailed is the enumeration of the species), and the ability to coordinate terms from different facets. These are now considered separately in terms of UDC.

4. Enumeration within facets in UDC

This takes two forms:

4.1 Continuous 'dissection' (Ranganathan's term for it) by 'irreversible' subordination–i.e., the different characteristics of division are applied in an order which would seem to be a 'natural' one each subordinate step being 'dependent' on the one above. The classic example would be a botanical or zoological chain:

596	Vertebrata
599	Mammalia
599.6	Ungulata
599.61	Proboscidea
599.614	Elephantidae

Similarly, in Building technology:

69.028	Openings
.2	Windows
.21	Hinged casements
.215	Side hung
.215.3	Opening outwards

or

696	Services, installations
696.4	Hot water service
696.46	Heat storage
696.463	Indirect system with calorifiers
696.463.5	Steam-to-water calorifiers

4.2 Separation into 2 or more subfacets (arrays), which can be compounded if necessary; for example, in Architecture, in the *Whole building* facet, application of the principle of division 'Function' would give a subclass *Residential* buildings; application of further principles (Single or multiple occupancy; Degree of detachment; Number of floors) would give a chain such as:

728	Residential buildings
728.3	Houses
728.31	Terraced
728.31.011.262	2-story

Here, although one might say the first two steps (Houses-terraced) are more-or-less irreversible (and the subfacets reflected not worth scheduling separately) the principle of Number of storys *could* be applied before 'function', to give, say

72.011.267:728.22	High rise buildings—Flats

4.3 In both cases, to some extent, and entirely in the first case, variation in detail depends on the cut-off point in enumerating the hierarchy. For example, in the 3rd ed. of the English Abridgement (1961), the first hierarchy is terminated at 599.61, the second at 69.028.2, and the third at 696.46.

4.4 Insofar as UDC notation is hierarchial, the differences in detail can be expressed roughly in terms of length of notation. There is no case to be made, however, for ruling a fixed number of digits since this would produce absurd variations in the size of documentary class produced (e.g., the 5 digits of 696.46 represent a far smaller class than, say 621.31 Electrical power supply, distribution and control—to say nothing of 621.38 Electronics.

4.5 The situation in the fourth example, in which one of the characteristics defining a step of division in the generic hierarchy (number of storys) is separately scheduled makes it clear that the detail within a facet is not only a matter of the degree to which any one hierarchy is

developed but depends also on the degree to which the different arrays within the facet are recognized. The theoretical distinction here is a rather tenuous one. For example, to terminate the development of the 'Houses' hierarchy at 728.3 would remove the possibility of distinguishing *Terraced* houses as a distinct species when searching, and it would do this by not recognizing the array 'Houses by degree of detachment'. In principle, this is no different from removing the possibility of distinguishing '2-story' houses by not recognizing the array 'Buildings by number of storys'. In practical terms however, the reduction in detail would be greater in the second case because the array 'Buildings by number of storys' is applicable to many kinds of buildings (i.e., *functional* kinds, like Residential, shops) whereas the array 'By degree of detachment' is really peculiar to *Houses*. The position here is, in fact, the appearance *within* a facet of the facility to form compound classes by synthesis (the notational face of logical intersection) referred to in 3.5 as the second facility needed to achieve resolving power.

5. Coordination by synthesis in UDC

5.1 There are 3 levels at which non-mutually-exclusive can be coordinated:

5.2 Coordination of terms from different subfacets (arrays) of the same facet; e.g., Houses–Single story 728.3.011.261 coordinates terms from the *Functional types* and the *Number of floors* subfacets in the *Whole building* facet of Architecture.

5.3 Coordination of terms from different facets; e.g., Painting–Ceilings 698.12.025.4 coordinates terms from the *Operations* and *Parts of building* facets of Building technology.

5.4 Coordination of terms from different main' classes ('phase relations', 'complex classes'). For example, Planning–land-use-Law 711.14:34.

5.5 In a fully faceted classification these 3 'levels' of coordination are separately provided for, since the resulting compound or complex classes should have a definite, consistent filing position in relation to each other. UDC doesn't provide consistently or distinctively for these different situations; e.g., the class. Thermal insulation of buildings by aluminum foil reflects strictly a (5.3) situation, but is represented in UDC as 699.86:691.771-416 as though it were a (5.4) situation. This does not matter so far as the achievement of precise index description goes. The fact that the colon in UDC can serve to represent any of the 3 modes of coordination (i.e., can be used as a general purpose synthetic device) means that theoretically (and, by and large, in practice) UDC has unlimited ability to define classes by coordination.

5.6 This ability may, of course, exact a heavy price notationally; e.g., the class Houses–Multistory–Kitchens–Regulations–Scotland would be 728.2.011.27.05:643.3(094.7)(411). Theoretically, if the notational consequences were disregarded, the maximum use of synthesis could specify a great deal of detail which we have assumed is obtainable only by enumeration within hierarchies. Taken to an extreme we might represent, say, the concept Internal combustion-engine by drawing on a common facet for the notion Internal (=191). In this way the enumeration in UDC could be reduced to a level approaching the 'minimum vocabularies' which some British users of coordinate indexing favour. However, UDC was never designed for this sort of use and we must assume enumeration as being the major source of detail within facets.

6. From the above analysis, it seems clear that the variations in resolving power of UDC between Full, Medium and Abridged editions is entirely a matter of the degree of enumeration of terms within hierarchies since the narrowing of classes by coordination (intersection) is *always* possible. Theoretically, this means that a word-count of enumerated terms would be an accurate measure of the differences in detail between editions. However, a complicating factor here is that UDC is not an entirely faceted system and many examples exist of enumerated classes which simply duplicate what is already achievable by synthesis. For example 631.3 Agricultural machinery yields a Special auxiliary =13 applicable to any agricultural produce (e.g. 633.1-13 Machinery in grain farming). But

at 637 Dairy produce one finds 637.12 Milk production and 637.125 Milking machinery. Again, it is well known that concepts represented by the Point of View numbers are extensively duplicated by enumeration in many classes.

7. Methods of arriving at different reduction in enumeration

There seem to be 2 different basic procedures possible:

7.1 Reducing the enumeration of subclasses within subfacets whilst retaining all these subfacets. For example, the recently published full edition of 656/656.7 Transport services, Traffic organization and control (BS 1000[656/656.7]:1968) provides this enumeration under 656.07 Administration and Management:

656.072	Passenger Services
-05	Persons (by Age, Race, Sex, etc.—as 3-05)
.1/.2	Activities of passengers (booking, boarding, etc.)
.3	Distribution of passengers
.4	Classification of passengers (by Class of travel, Number traveling, Activity (tourist, etc.)
.5/.7	Administrative operations on or for passengers (inspection, care, catering, etc.)

Most of these subfacets have a fair number of classes in them, as indicated in my parentheses. The first stage of reduction (for a Medium edition?) could eliminate the enumerated classes within subfacets whilst retaining the subfacet as a *class*; e.g., a document on *family* travelling would get the number 656.072.4 which would represent all documents dealing with particular classes of passengers (defined by the special conditions of travelling rather than by the more generally applicable principles covered by −05).

7.2 Reducing the enumeration by eliminating not only of the classes within arrays (subfacets) but the arrays themselves. In the above example, this would leave only 565.072 Passenger services. This is, in fact, what the 1961 Abridged English edition does.

7.3 This example demonstrates again the important part played by synthesis in providing detailed specification in UDC. The existence of the common facet for *Persons* at 3-05 means that even in the drastic abridgement to 656.072 the notion of *classes of passengers* can be specified (656.072:-05) and a fair number of subclasses within this facet. This reflects a general situation whereby specificity is absolutely reduced only for those hierarchies which depend solely on enumeration and cannot utilize synthesis with other facets because they represent classes uniquely dependent on the given context (as, for example, passenger activities like booking, boarding, etc. are peculiar to the context of Transport services).

8. We are still left with the problem of how far to go in the full editions and where to draw the cutoff line in reducing hierarchies. All that can be said, it seems, is that both decisions should reflect literary warrant in the sort of documentary material the edition is designed to organize.

8.1 The commonest measure of literary warrant for a general collection is the output of monograph literature represented by large current national bibliographies. If we use BNB (British National Bibliography) as a guide here, this would give us a level of specificity a little greater than that catered for by the Dewey DC. Since the function of the Abridged edition is to serve the classification of 'non-core' material in special libraries this would seem a reasonably appropriate level for it to aim at.

8.2 For the Full edition there is still something of a mystery. Lloyd, at FID, argues that extreme detail (of the kind, presumably, demonstrated in those parts of the Seismology schedule excluded by FID) is inappropriate in an index language for pre-coordinate indexes. It should be obtained, he says, from descriptors in a thesaurus. This seems to me to underestimate the use of UDC as a thesaurus itself. Thesaurus terms divorced from a classificatory context which defines them and indicates their major relations to other terms are diminished in their usefulness. Also, the summarization principle which is basic to precoordinate indexing often calls for the recognition of these super-specific terms and I think a Full edition should cater for them too.

8.3 The use of computer printouts showing the frequencies of postings to individual terms is used by a number of post-coordinate indexing systems to control the size of their vocabularies. This may have something to offer to UDC editions, but I'm not sure that the function of UDC as a *general* index language, from which specialist users will select what they want and reject what they don't want, does not make this criterion invalid.

FACETED ANALYSIS WITH THE UDC (T.W. Caless)

The universe of discourse evolves from what is discovered by man about every atom and molecule that exists and of all sets and combinations of these atoms and molecules, which comprise an organism at any time. When organisms reproduce, this activity is observed and recorded by man as information about that activity. This information is further complicated by man, the observer and recorder, reacting to his own feelings and thoughts plus a reaction to being acted upon by his environment. Information is, therefore, a recorded by-product of scholarly and non-scholarly investigation which undergoes a series of complex interactions.

The development of a finding language requires that we draw upon the natural language for the purpose of locating a sought for piece of recorded information which will contribute to our understanding of a particular investigation. This language is abbreviated in that it is not used for speaking or writing but is composed of fundamental subject elements specified by constituent relations which may be further linked with logical relations.

An information language, to be fully effective, must be developed so that the user can find things naturally. At the same time, however, it must be applied to the literature consistently or it will not be possible for the user to predict where he might locate a specific piece of information if it were arbitrarily filed. Once a filing system has been developed, it should represent the best overall arrangement for that discipline(s) even though individual subject specialists might occasionally find reasons to quarrel with it.

If we recognize that the element of predictivity is important in the search, then procedures for linearization of concepts or categories of information for the disciplines included in the collection need to be developed. A prescribed sequence allows the necessary translation of the terms of an author into a recognizable pattern, however loosely they might have been expressed, and the same rules can later be applied to translate the terms of an inquiry into the same pattern to produce a match.

Developing procedures for linearizing categories for a discipline requires that we identify the most concrete *thing* that we investigate in a discipline, its *kinds*, *materials*, *processes*, *properties*, *operations*, *agents*, *viewpoint*, *time* and *form*. If there are kinds of kinds, or materials, etc., then they should also be identified in this analysis. Using three major topics which are instrumental in human development, we note that we have:

Physical
Society
Self

which represent the most concrete *things* that we investigate. These main classes may be further expanded as:

Physical Organism
Society
  Peer Groups
  Cultural
  Affection
Self
  Development
  Adjustment
Psychic Phenomena

Note that by adding the last category of Psychic Phenomena, we now have a general to special arrangement of classes as we read down the list. This may also be likened to an arrangement which goes from a Static condition (physical) to a Dynamic condition (society, self, and psychic phenomena). By following the linear string of concepts mentioned earlier, our document analysis will proceed as follows:

Example 1

DOCUMENT: Dewey, John. "The child and the curriculum", Chicago, Univ. of Chicago, 1902, 31 p.

ANALYSIS:
Thing–Psychic Phenomena Viewpoint
Kinds–Child
Material
Processes
Properties
Operations
Agents–Curriculum
Time–1902
Form

INDEXING ENTRY: CHILD: CURRICULUM: 1902

In this example, the reference to this document would then be found in the file under the category Psychic Phenomena. The other categories remain empty for this analysis.

Example 2

DOCUMENT: Isaacs, Susan Sutherland (Fairhurst). "Intellectual growth in young children", London, Routledge, 1930. 370 p.

ANALYSIS:
Thing–Psychic Phenomena Viewpoint
Kinds–Child, young
Materials
Processes–Intellectual Growth
Properties
Operations
Agents
Time–1930
Form

INDEXING ENTRY: CHILD, YOUNG: INTELLECTUAL GROWTH: 1930

The results of this analysis are similar to Example 1 and the reference to this document will also be located under the category Psychic Phenomena. Note that the subordinate term "intellectual growth" could have been entered as "growth, intellectual", but these are options that subject specialists will need to agree upon as the system is developed.

Example 3

DOCUMENT: Illinois, University of. Laboratory of Personality Assessment and Group Behavior. "Prediction and understanding of the effect of children's interest upon school performance; a correlation study of ability, personality and motivation factor measures in relation to criteria of achievement." Urbana, 1962. 77 p.

ANALYSIS 1: Thing–Self Development
Kinds–School Performance
Materials
Processes
Properties
Operations
Agents
Viewpoint
Time–1962
Form

INDEXING ENTRY: SCHOOL PERFORMANCE: 1962

Example 3 is typically much more complicated than Examples 1 and 2. Our first analysis revealed that this document would be identified as appropriate to the category Self Development. This single analysis does not serve adequately in summarizing all of the concepts which are present in this document, since we note that the subjects which are present additionally cross into another major category. Therefore, further analysis is necessary.

ANALYSIS 2: Thing–Psychic Phenomena
Kinds–Child Interest
Materials
Processes
Properties
Operations–Prediction and Understanding
Agents
Viewpoint
Time–1962
Form

INDEXING ENTRIES: CHILD INTEREST: PREDICTION: 1962
CHILD INTEREST: UNDERSTANDING: 1962

ANALYSIS 3: Thing–Psychic Phenomena
Kinds–Ability, Personality and Motivation
Materials
Processes
Properties
Operations–Achievement Criteria
Agents–Statistical Correlations
Viewpoint
Time–1962
Form

INDEXING ENTRIES: ABILITY: ACHIEVEMENT CRITERIA: STATISTICAL CORRELATIONS: 1962

PERSONALITY: ACHIEVEMENT CRITERIA: STATISTICAL CORRELATIONS: 1962

MOTIVATION: ACHIEVEMENT CRITERIA: STATISTICAL CORRELATIONS: 1962

This document now may be referenced or located under two major categories with a total of six access points. This demonstrates that the only limit to the number of analyses required for adequate summarization is dictated by the nature of the subjects. Note also that the linear ordering of the concepts does not reduce the flexibility of the technique but the necessary controls for retrieval are always present.

Additional manipulation of the concepts within each indexing entry string shown above is possible, depending upon requirements. It is possible that Time, for example, would more effectively be used for chronologically arranging Case Records. If this is agreed upon, then the file for Case Records would be partitioned by year and an indexing entry would appear as:

1962: Psychic Phenomena: Motivation

and the entry would be interfiled in Category 1962 and be alphabetically sequenced in the file. If this is done consistently, then retrieval is possible. The user is the most important aspect of an information system, and his needs must be provided for in the overall design of any system. There are numerous other configurations possible with the linear pattern described here, and once there is agreement to a particular design, it should be tried and demonstrated before implementation. Notational signs in addition to the colon are also frequently used as a separator between concepts in indexing such as the comma, parentheses, and dash. Regardless of the notation selected, entries within the file can be sequenced alphabetically by letter characters and/or numerically, depending upon the indexing entries.

The manual search is quite elementary and does not need to be expanded upon here. The computer search will require understanding of the indexing technique. Additional familiarity with the use of Boolean Operators will be of assistance. For example, by using the previously cited indexing examples, let us assume that our requester asks for documents on Intellectual Growth in Children. We would know immediately that we would need to search the file on Psychic Phenomena for CHILD: INTELLECTUAL GROWTH. Note that the document in Example 2 would *not* be retrieved unless we additionally specified CHILD, YOUNG: INTELLECTUAL GROWTH. If the requester agreed that both of these entries were important to his request, then the file on Psychic Phenomena would simply need to be searched twice. Or, if a requester wanted CHILD: INTELLECTUAL GROWTH to include CHILD, YOUNG: INTELLECTUAL GROWTH: CHILD, SMALL: INTELLECTUAL GROWTH, etc., the file format can be prepared and the computer programmed to truncate all concepts which follow CHILD (with or without the comma) but precede the colon and accept as a match to a request CHILD: INTELLECTUAL GROWTH. This will broaden the search considerably, and will save searching the file one or more times if a general search is required.

Developing an information system is self-organizing and after the indexing staff has analyzed several hundred documents, it will be found that many documents will fall readily into place if the above scheme is followed. This does not mean that all will, however. Authors can be quite vague and subjects of documents can be elusive. A good policy to follow in analysis for indexing is not to second guess the author's intent by bringing non-subject concepts into the analysis. Otherwise it follows that the retrieval will locate material which is not related to the request. The analysis must be directed towards identifying the subject of the document, regardless of how disguised it may be. As was mentioned, this system will be self-organizing and after several months of effort, the indexing terms used will need to be examined, the synonyms resolved, and the mapping of the terms into the partitions of the linear indexing string or citation order accomplished. In fact, this will need to be done periodically through the years as a means to controlling the expanding vocabulary.

The pattern of analysis followed above is applicable to any subject. An example of the subject of Oceanology (the science of the salt water hydrosphere) divided into categories of terms (facets) from an actual analysis of over 1,000 documents is as follows:

THING/KINDS - Bays, Gulfs, Oceans, Seas, etc.
PARTS - Depth, Mid-Depth, Surface, Layers, etc.
CONSTITUENTS - Copper, Iron, Nitrate, Oxygen, etc.
PROPERTIES - Alkalinity, Density, Salinity, etc.
PROCESSES - Cooling, Currents, Storm Surges, Waves, etc.
OPERATIONS - Cruises, Expeditions, Observations, Forecasting, etc.
AGENTS - Buoys, Drift Bottles, Thermometers, Unmanned Stations, etc.

Of course, additional diffuse (with respect to Oceanology) facets were identified in the analysis which may be likened to the UDC Common Auxiliaries. A similar analysis of documents in Meteorology and Seismology was performed and categories of terms identified for those disciplines. It is apparent that the most substantive facet (THING/KINDS of THING/ACTIVITY) is the primary facet of those disciplines and PARTS through to AGENTS are subsidiary facets which can be interconnected for class specification in the facet order shown. It is quite likely that continued analysis of these topics will reveal additional subdivisions of SHAPE of OBJECT OF ACTION and these will need to be merged into the above

arrangement for completeness. Investigators of faceted classification schemes have also felt that some sort of relational devices need to be developed for better class specification and Farradane[1] and Perreault[2] have offered some untested solutions to this problem. It is interesting to note that even though Research Scientists will continue to discover new information about these disciplines of the Earth Sciences, it is unlikely that the categories and their overall order will be affected; only new terms within the categories will need to be added.

It is this analysis approach which will provide the basic indexing vocabulary, eventually leading towards a revised UDC schedule. Once the vocabulary is properly structured and tested, it is then a simple matter to add the UDC notation. And by following these procedures, the UDC schedules will then guide the analysis of future documents received for our collection.

It has been found that arranging the categories THING, KINDS, PARTS, CONSTITUENTS, etc., across the top of an analysis sheet with successive analyses proceeding vertically down the paper (similar to the multiple analyses required in Example 3 for a Human Development document), this visual display of order will serve as an aid to the Analyst. By visually displaying the facet formula, inconsistencies in the analysis can be reduced. This approach has been further discussed by Caless[3-4] and Perrault (see page 16).

A PROPOSED FACET FORMULA FOR UDC 551.5 (METEOROLOGY) (J.M. Perreault)

In order to treat a class from a more-or-less enumerative classification such as UDC as a faceted class capable of synthesis of complexes in a predictable way, it is necessary to introduce a facet formula which can guide both the indexing and the searching. UDC does not have such explicit facet-formulary rules, though they are often implicitly present. The attempt made below is to spot the various facets contained in UDC 551.5, to determine their optimum order as applied to the statement of documentary contents, and to predict the utility of such a facet formula in indexing and searching. The attempt is carried out by allocating all numbers in UDC 551.5 to one of several columns in a matrix, and by stipulating that the contents of this matrix must be cited in the complex number in the order given.

The partitioning is carried out only at a relatively gross level; there is still much to be done in terms of examination of each number to see whether it would not be better relocated in a matrix column apart from that indicated for its superordinate numbers. For instance, would 551.577.6 not be better placed out of its [P_2] column into the [A] (see p.18) column?

The plan of attack is (a) to see whether the whole class UDC 551.5 can be analyzed in terms of Ranganathan's categories (first round); (b) to see whether levels and further rounds can successfully accommodate all concepts in the field not covered in the first round; and (c) to point out several unresolved problems and some possible solutions to them but without attempting to solve all problems of misplaced concepts and/or inseparable complexes which resist facet analysis.

Analysis Into Gross Categories

The discipline of meteorology seems to be divided, for purposes of classificatory analysis and translation, into the gross categories of Phenomena, Energy, Space, Time and Form. These match the UDC codes:

551.51/.7 [P]
1/551.4, 551.509, 551.59/9 [E]
(1/9) [S]
"0/7" [T]
551.506, (02/09), -0/-9 [F]

Thus a document representing long-term observations, 551.506.3, of thunderstorms, 551.515.4, in 19th Century, "18", Scotland, (41), would be combined in the order implied in the matrix:

[P]	[E]	[S]	[T]	[F]
551.51/.7	1/551.4, 551.509, 551.59/9	(1/9)	"0/7"	551.506, (02/09), =0/=9
551.515.4		(41)	"18"	551.506.3

A relational sign, of course, is necessary to join main-class numbers, giving 551.515.4(41)"18":551.506.3 as the final result.

Another document might be a bibliography, 016, of winter, "324", gales, 551.553.8, which would plug into the matrix as:

[P]	[E]	[S]	[T]	[F]
551.553.8			"324"	016

Again a relational sign is needed to build the final code, 551.553.8"324":016.

As a final example at this gross level of analysis let us take a document on lunar influences, 551.590.22, on fronts, 551.515.8, namely a set of weekly bulletins, 551.506.1:

[P]	[E]	[S]	[T]	[F]
551.515.8	551.590.22			551.506.1

All three numbers need colonning here, giving 551.515.8:551.590.22:551.506.1.

Analysis of Gross Categories Into Sub-Categories

There are, as well as the gross categories mentioned, several levels and rounds noted within the meteorological literature, and fairly well represented in UDC. Phenomena are either concrete events representing a complex of characteristics, 551.515, or simpler sets of characteristics such as 551.510/.513, 551.52/.57.

These are to be arranged by decreasing concreteness thus:

[P1]	[P2]	[P3]	[P4]	[P5]	[P6]
(Complex phenomena)	(Constituents)	(Wind)	(Pressure)	(Temperature)	(Mechanical properties)
551.515	551.57	551.55	551.54	551.52	551.510/.513

A document on rain, 551.578.1, resulting from thunderstorms, 551.515.4, would plug into the matrix as:

[P1]	[P2]	[P3]	[P4]	[P5]	[P6]
551.515.4	551.578.1				

one on impurities, 551.510.42, involved in rain, 551.578.1, as:

	551.578.1				551.510.42

one on the effect of frozen soil, 551.525.5, on the barometric gradient, 551.542.1, as:

551.542.1 551.525.2

Among the several codes shown above as [E], the best citation order is probably 551.59[E_1], followed by all outside codes as [E_2]. However, it may be the case that this order is more a precedence order than a citation order, just as that among the various levels of [P], where it is highly unlikely that more than two levels could occur as mutually modifying elements in a single document.

The best order for [F] is special forms, 551.506[F_1], followed by general form, (02/9)[F_2], and then by language =0/=7[F_3]. A complicating factor is the lack of (01) as a general form, and the need to colon on main-class 01- codes as substitutes.

Now we have a gross order of categories, [P, E, S, T, F], and a more developed form of it, [P_1, P_2, P_3, P_4, P_5, P_6, E_1, E_2, S, T, F_1, F_2, F_3]; more or less standard sub-facets of [S] and [T] can also be developed. But there are parts of the 551.5 schedules not accounted for in this citation order, and they seem to fall into two sectors: (a) a second round, to be introduced after [E_2], consisting of [$_2P_1$] instruments, 551.508, [$_2P_2$] vehicles, 551.507, [$_2E$] methods, 551.501, [$_2S$], and [$_2T$]; and (b) the idea of application or results, represented by 1/551.4, 551.58, and 551.7/.9. The second of these, being postulated as target (with meteorology itself as the source) necessarily precedes all meteorology codes in the citation order; labelling this one as [A] for application, we get an over-all citation order, [A, P_1, P_2, P_3, P_4, P_5, P_6, E_1, E_2, $_2P_2$, $_2P_2$, $_2E$, $_2S$. . . , $_2T$. . . , S. . ., T. . ., F_1, F_2, F_3].

Some Unresolved Problems

Within each column of the resultant matrix can still be seen a lack of total logical mutual exclusion: for instance, until recently 551.508 (instruments) had two sub-sub-facets: (a) 551.508.1 (instruments for upper-air investigation) and (b) 551.508.2/.9 (instruments for measuring radiation, temperature, pressure, humidity, and mechanical phenomena); 551.509 now has two sub-sub-facets; one for forecasting and one for modifications. There are also such disquieting features to be noted as a similarity of pattern in the break-down of such classes as 551.524.1/.36, 551.511/.513.6, 551.553.1/.5, 551.571.1/.36, 551.577.1/.36, all of which could better be reduced to a special auxiliary; and a borrowing-feature for the divisions of 551.515.1 which can be applied to 551.515.2, .3, .4, and .7; but these divisions (e.g. Tracks: 551.515.13 of barometric depressions, 551.515.23 of hurricanes, 551.515.33 of tornadoes, 551.515.43 of thunderstorms, and 551.515.73 of anticyclones) cannot be exhibited alone, i.e. in general.

Similarly, it is to be regretted that a great deal of redundancy will inevitably occur with such a code as 551.515.8:551.990.22:551.506.1; the form-class 551.506 might conceivably be reduced to a -1/-9 auxiliary, and the classes 551.501, 551.507, 551.508, and 551.509 to a .01/.09 auxiliary. Then such a code could be economized to read 551.515.8:551.590.22-1; 551.515.4(41)"18":551.506.3 to 551.515.4(41)"18"-3; a real horror like 551.515:551.509.1:551.508.5:551.507.351:567.501.3:551.506.1 (weekly bulletins on Beaufort scale observations of formations; the instrumentation is an anemometer brought into place by a kite, and the method of transmission is radio) to 551.515.091.033.073.51.013-1.

The final (and probably insurmountable) obstacle is the misplacement of concepts; an example is 551.577.5/.6: .5 should be coloumed on from 551.59, .6 from 551.58.

ORDER OF OPERATIONS AND USE OF SQUARE BRACKETS (J.M. Perreault)

For employment of a pre-coordinated indexing language with a computer (or, indeed, for efficient clerical manipulation of it it is necessary for certain syntactical rules to be followed. Otherwise interpretation of the complexes of terms, each of which complex is postulated to represent a document (or, to use Swerget's felicitous expression, a "documentary unity" whether encyclopedia, treatise, chapter, article, section, paragraph, or sentence) will be rendered mere guesswork.

Citation order, insofar as it is dictated not by mere rigor (either for the sake of predictability or of convenience), but by the exigencies of meaning, is a means to the elimination of guesswork in interpreting pre-coordinated complexes of terms. But the simultaneous advantage and disadvantage of a synthetic classification such as UDC is that occasions will arise for which not even the most carefully prepared citation-order rules will be sufficient. To aid in the formulation of guesswork-free complex index-statements will require meaning-empty symbols capable of showing the boundaries of syntactic relationships and of semantic influence, much like punctuation in natural languages. (It must be noted that in UDC the punctuationals are often more than merely syntactic, they are more normally semantic primarily and syntactic only accidentally, as when parenthesized expressions are intercalated to change a convenience-dictated citation order.)

An ineluctable question in a computer-manipulated pre-coordinated indexing language is: Is there the possibility of expressions of a higher order of complexity than the linear sort expressed as A > B > C, where the main subject A is modified by B, and the resulting complex A > B is modified by C? If B represents a place-modification, of A, by B *and* D, the whole complex/compound being modified by the form C, will A > B > D > C be our only (and incorrect) available codification?

Or, to put the question in more general terms, as the computer reads through the digits of the complex, can it do anything more than just to treat the whole string like a long word—or can it treat it like a sentence, with clauses, phrases, etc.? (The question of modes, tenses, *etc.* is not asked—perhaps it would make no sense here; we may speculate that the sole predicational form in information work is the categorical.)

Given any expression in UDC that goes beyond mere scheduled elaboration, intelligibility demands that there be a rule for the assimilation of each successive digit. For instance, in the expression 550.342(747), the decimal point conveys nothing, so that after the first three operations 550, the fourth is not ⌴, but ⌴3. After the sixth operation, again, the seventh is not ⌴(, since again, there is no real meaning to (by itself. Therefore the whole set of operations runs 550.342(747); that is, when a symbol is encountered which initiates a term of the complex, that whole term is read in before the whole term is assimilated to the preceding whole term.

But this is a minor point compared to the problems occasioned by a complex in which more than two terms are joined. The ordinary rule A > B > C would apply to a complex such as 550.342(747)(05), with each major term being represented by one of the three letters, as well as within each term. But in a complex in which either *a*: the same initiating symbol occurs more than once, or *b*: the various initiating symbols occur in an order different from the reverse of that shown on p. 10 of 3d abr. English edition. What about an algebraic order such as A ⌴B ⌴C instead of the normal A ⌴B ⌴C? For such an overriding to be intelligible, there must be a normal order for it to override. From the examination of 550.342(747)(05) it will be intuitively agreed that each new digit 0 1 2 3 . . . . 9 is to be assimilated to whatever precedes it, without a new "word" starting; that once a new word starts with its characteristic initial symbol, further digits are assimilated until the terminal symbol is read; that certain initial symbols have greater "bond strength" than others.

A provisional tabulation of an order in which each type of juncture takes place in UDC was attempted in my paper "Towards Explication of the Rules of Formation in UDC"; discussion between Antony Foskett, Derek Langridge, Jack Mills, and myself, in the course of consultations directed by Thomas Caless, led to an agreement on a rather different order, namely (following after the ordinary numerical digits):

n'n	*n/n*	*n+n*	.*On* *-n*	*n::n* .*00n*	(=*n*) (*n*) "*n*"	*n:n*	*(On)*	=*n*

This means that in A'B/C A and B would be joined before B and C, as A'B/C while A/B'C would be A/B'C. Symbols arranged between braces are simultaneous in operation (equivalent in bond strength); thus it is their citation order that determines order of assimilation, according to the standard A > B > C rule. This gives A.OB-C, but A-B.OC, *etc.* Intercalation of self-enclosed terms such as (=*n*), (*n*), "*n*", and (*On*) is done with the normal technic: standard order 338(44)"18" can become 3(44)38"18", or 33"18"(44), or even (by double intercalation) 3(4"18"4)38. But intercalation of non-self-enclosed terms encounters the difficulty that the terminus of the term is difficult or even impossible to recognize; the square brackets are drawn upon to show the overriding misplacement, without themselves substituting for any other symbol (therefore n o t in agreement with 3d English abr. p.11). Discussion among the group resulted in agreement that square brackets must have e i t h e r *a*: symbol inside, in which case they are interpreted as being used to intercalate, or *b*: a symbol outside, in which case they are interpreted as being used for algebraic grouping. The first case would be such as 620.1[:669.14]74 (note that 3d English abr. p. omits the colon); the second such as 16:[17:7], where lack of square brackets would give the order of assimilation A B C instead of the desired A B C.

The situation until recently was that the square brackets were used for three distinct purposes without any way of telling the one from the other. These three were Intercalation, Algebraic grouping, and Subordination. The third of these burdens has been more or less removed from the square brackets by adoption of the double colon *n::n* (though there are unresolved problems not to be tackled here); the use of a symbol such as the colon either inside or outside the square brackets themselves effects the necessary distinction.

P-note version: (distributed to the CCC by the FID in November 1969 as C69-27)
It should become a binding and universal convention that

(a) Any UDC code consisting of more than two elements is normally interpreted as assimilating each such element from left to right (except as provided below). *E.g.*, 338(44)"18" or 75.023.2:667.6:669.71—each interpreted as A B C.

(b) Any UDC code consisting of more than two elements, with junctures formed by two or more different relational symbols or auxiliaries, is interpreted as assimilating each element in an order fixed by the following tabulation: (numbers in parentheses above each UDC combination signify the combining precedence relative to the others):

(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
[ ]	*n'n*	*n/n*	*n+n*	.*On* *-n*	*n*::*n* .*00n*	(=*n*) (*n*) "*n*"	*n*:*n*	*(On)*	=*n*

E.g., 661:546.33'12, interpreted as A B C, not A B C

05+07:32+8, interpreted as A B C D, not as A B C D

378.4:820"1837/1901", interpreted as A B C

378.4:820(05), interpreted as A B C

(c) Any UDC code consisting of more than two elements may be forced to vary from the order of assimilation described above in (a) and (b) by the use of square brackets, which are purged of any independent meaning. Thus the example 620.1[669.14]74 (English abr. 3d ed., p.11) is rendered invalid; the normal form, 620.174:669.14, implies that the colon needs to remain (since the square brackets have no independent meaning) in 620.1[:669.14]74. This variance from ordinary assimilation-conventions (as well as from standard citation order) is that commonly called intercalation, and always requires retention of the moved relational sign, whether *n+n*, *n/n*, *n:n*, or *n'n* inside the square brackets.

(d) Any UDC code consisting of more than two elements may have internal groupings which need display in order that the order of assimilation described above in (a) and (b) do not fragment them, thus changing their meaning. The square brackets will be used to show the boundaries of such internal groups of elements, thus overriding the (a)- and (b)- conventions. *E.g.*

[05:32]+[07:81](72), interpreted as A B C D E,

not (as without the square brackets) as A B C D E

In contradistinction to (c), this convention always requires retention of a relation sign, whether *n+n*, *n/n*, *n:n*, or (though unlikely) *n'n* o u t s i d e the square brackets.

CLASSIFICATION OF SCIENCE AND TECHNOLOGY (A.C. Foskett)

There are various ways of arranging the topics which fall within the broad heading of Science and Technology. The UDC follows Dewey's lead in gathering together all of Science (with the exception of medical science) in one group, and all of Technology in another. At the other extreme is Brown's practice in the Subject Classification of collocating each technology with its basic science; this idea has often been ignored, largely because Brown's execution of it left much to be desired, but is seen in Colon Classification and in Bliss's Bibliographic Classification. Ranganathan follows Physics by Engineering, Chemistry by Chemical Technology, Geology by Mining, Zoology by Animal Husbandry, etc. Bliss collocates Chemistry and Chemical Technology, and *some* branches of Physics with their technologies--but then leaves the rest to go into his Useful Arts class towards the far end of the overall sequence. Colon also has a Useful Arts class which is something of a 'dustbin' collection.

It seems clear that there is no general consensus among the schemes. Literary warrant works both ways: one finds textbooks in which the various branches of Physics are brought together, but one also finds such works as the 'Sourcebook in Atomic Energy' which deals with both Nuclear Science and Nuclear Technology.

It is equally clear that there is no general consensus among the individuals concerned. In the main, a mechanical engineer will have more in common with an electrical engineer than with a physicist interested in Mechanics; an electrical engineer more in common with a chemical engineer than with a physicist interested in Electricity and Magnetism. This is however by no means always the case; a vacuum physicist and a vacuum engineer will have much in common, as will a solid state physicist and an electronics engineer working with transistors or integrated circuits.

In UDC the use of the + may provide a way out of the dilemma within the existing framework, by enabling us to specify both the science and the technology of a particular topic, e.g.,

539.1 + 621.039	Nuclear science and technology
533.5 + 621.52	Vacuum science and technology

This is the use of the + to indicate a genuine aggregate class, but it is open to some theoretical objections. These can best be demonstrated by considering the overall order resulting from its use.

The arrangement in Figure 1, which is that arising from the straightforward use of the plus +, obviously gives an overall order in:

5/6	Science and technology
5	Science
5...	...
53	Physics
...	...
533.5 + 621.52	Vacuum science and technology
533.5	Vacuum physics
....	
539.1 + 621.039	Nuclear science and technology
539.1	Nuclear physics
6	Technology
...	...
62	Engineering
...	...
621.039	Nuclear engineering
...	
621.52	Vacuum engineering

Figure 1

where Science comes between the general heading Science and Technology and the special-general headings Vacuum science and technology and Nuclear science and technology. On the other hand, the existing UDC notation does not lend itself to the arrangement given in Figure 2, which does give the desired progression from general to special. To achieve a mnemonic effect, it would be necessary to have a parallel arrangement under the general heading Science and Technology and under the special headings Science *and* Technology; without this, it would be difficult to provide in advance for every possible literary warrant at the wider heading without a great deal of enumeration. This however implies that there *is* such a parallelism between Science and its related Technologies. I am not convinced that this is the case; even Colon Classification shows little parallelism *within* its parallel main classes, and other schemes show even less.

5/6	Science and technology
?	Nuclear science and technology
?	Vacuum science and technology
...	...
5	Science
...	...
53	Physics
...	...
533.5	Vacuum physics
...	...

539.1	Nuclear physics
6	Technology
...	...
62	Engineering
...	...
621.039	Nuclear engineering
...	...
621.52	Vacuum engineering

Figure 2

It may be that an approach using the Classification Research Group (CRG) analysis of entities and attributes[5], combined with integrative levels, would give a more satisfactory comparison, but I am inclined to doubt this.

I am in fact not at all sure that any solution which is both theoretically and practically satisfying exists, but it may be possible to use some device to give a satisfactory result with UDC. The use of [...] suggests itself as a possibility: the question marks in Figure 2 might be replaced by notation of this kind, e.g.,

5/6[533.5 + 621.52]
5/6[539.1 + 621.039]

The only objections I see remaining to this are those which apply to the overall order within UDC; the divisions at 5/6 would parallel those at 53/54 (or possibly a wider set of the subdivisions of 5) and would therefore be open to the same criticisms. Since we have in mind the improvement (admittedly long-term) of this situation, perhaps we should go ahead and recommend the solution proposed here.

A third possibility might also be considered. Many works which require the use of the + in this way contain a minimum of the science - sufficient for the student to gain the foundations for a study of the technology. In such cases, it might well be desirable to file works on the science and technology of the subject immediately before works on the technology, e.g.,

621.039 + 539.1	Nuclear science and technology
621.039	Nuclear technology
....	
621.52 + 533.5	Vacuum science and technology
621.52	Vacuum technology

Any solution to this problem is likely to leave some users dissatisfied!

INSTRUCTIONS FOR UDC SCHEDULE REVISIONS (A.C. Foskett)

Preamble. The UDC has from its beginning used several synthetic devices in its notation to facilitate the construction of class numbers for composite subjects. These include the various common and special auxiliaries, and in particular the colon, which is widely used to signal "in relation to". UDC schedules have thus been 'synthetic' from their beginning. The basis on which UDC was constructed, Dewey's Decimal Classification, 5th edition, was however largely an 'enumerative' scheme; i.e., one in which composite subjects are listed as they stand rather than being constructed from their individual elements. UDC took over much of this enumerative structure, and this has in many cases led to problems where there is a clash between an enumerated place for a composite subject and also the possiblity of synthesising one. An example can be found in Chemical Technology - mineral oil processing 665.5. Here, fractional distillation is enumerated at 665.52, but the notation might equally well be 665.5.048.3, constructed by adding the

auxiliary .048.3 (meaning Fractional distillation) to the base number 665.5. The confusion is accentuated by the fact that only a few operations are enumerated at 665.5., but as many as possible are included in the auxiliary table; if we use the enumerated subdivisions, we shall end up with an order which is not satisfactory, for it will consist of a mixture of enumerated and synthetic subdivisions.

Problems also arise with the indiscriminate use of the colon. Relationships between subject may be of several kinds; if we only have one symbol, the colon, for all of these kinds of relationships, we are likely to find again that the resulting arrangement will be unhelpful and confused. The American Institute of Physics research into the use of UDC with mechanized systems showed that the lack of precision in the use of the colon was a significant drawback.

These and other problems may be solved by the use of the 'analyticosynthetic' approach. First used by Dewey, but only implicitly and in a limited fashion, this approach was put on a sound theoretical basis by Ranganathan, and has since been further developed by classificationists throughout the world. Many UDC schedules compiled over the past few years have demonstrated this method, and it is proposed that it should become the standard method of compiling new or revised schedules. To facilitate this, the following notes are intended as a guide to the classifier about to compile a schedule, indicating the kind of problem likely to be encountered and the possible solutions.

Procedure for Constructing a New UDC Schedule

1) The first essential is to make sure that the subject area being covered is homogeneous. An explicit definition of the subject should be agreed, and concepts falling outside the subject thus defined must be rigorously excluded, otherwise it will prove impossible to devise a satisfactory schedule. For example, when the schedule for Seismology was revised, the subject was defined as: the science of Earth Disturbances in all that relates to their forces, duration, lines of direction, periodicity and other characteristics. On the basis of this definition it was found necessary to remove Earthquake engineering to 622 and Earthquake disasters to 614, since neither of these topics belongs in Seismology, though both were previously to be found there. Similarly, in revising the schedule for Library Science one would need to remove those subjects which, through historical accident, are found there but do not belong there, such as Bibliographical psychology.

2) The subject should then be analyzed into its facets. A facet is a set of terms all of which bear the same broad relationship to the inclusive class; for example, in Agriculture one will find a *Crops* facet, in Literature a *Literary forms* facet, in Sociology a *Persons* facet. Facet analysis must be based in the first place on a study of collections of at least several hundred documents, but this is the position in which most potential revisers find themselves, in that their proposal stems from practical needs of their document collections. There will normally be only a limited number of facets in any given subject. Many of these will become obvious very early, others may occur only rarely.

It will usually be necessary to analyze a facet still further into a set of arrays or subfacets. For example, a Persons facet would need to be organized into arrays such as Persons by age, by sex, by ethnic group, by social status, by occupation, by nationality, by height, by weight, by color, and so on. The analysis should be carried out to the point where the terms in any given array are *mutually exclusive*: i.e., they cannot be combined in a composite subject. For example, child, adolescent and adult can appear in the same array, but child and male cannot, because there is the possibility of combining child and male to denote boy. Equally, analysis into child and male will exclude boy.

The analysis should not however go beyond the level of terminology used in the document collection. For example, if the documents refer to Thermometers, this is the term that should be used, rather than Instrument for the measurement of temperature. This analytical definition will in fact be reflected implicitly by locating the term Thermometer in the Agents facet applicable to the measurement of heat.

After some idea of the structure has been obtained by this study of the documents, the vocabulary should be enlarged by studying such reference tools as dictionaries, encyclopedias and glossaries, as well as existing classifications, thesauri, and lists of subject headings. If the analysis has been adequately carried out, the study of these will not lead to any increase in the number of facets or arrays, but it will certainly increase the number of terms to be included.

Most work on facet analysis has taken place in the area of Science and Technology, and here it is possible to indicate the kind of facet that is likely to be found, for example: Things, Kinds of things, Parts, Constituents, Properties, Processes, Operations, Agents. However, it should not be thought that these are the only kinds of facets likely to occur; in all cases, it is the literary warrant of the collection of documents that will dictate the structure of facets and arrays, and generalized statements need not impose any constraints. Rather, their use is to suggest the kind of pattern that may appear.

3) Once the structure of facets and arrays is complete it is possible to oragnize their contents into a helpful order. Within an array there may well be hierarchies of concepts related as genus to species; in such cases, the more specific should follow the general. Among concepts which are of equal rank, i.e., coordinate, there are various principles which may be helpful in giving some indication of useful arrangements. These include.

chronological
evolutionary
increasing complexity
spatial
size
alphabetical (for concepts having specific *names*)
traditional

The latter two should be used if no other more useful principle seems appropriate. It may also be useful to bear in mind the idea of 'preferred category' i.e., the idea that one particular concept is more significant than the rest and should therefore appear at the head of the list. For example, UDC considers Earth in Astronomy outside the normal sequence of planets, arranged according to their distance from the Sun.

The analysis carried out so far will have given a list of terms arranged in facets and arrays, showing their relationships within those arrays. Although it is not the primary purpose of the UDC, such a listing will form the basis of a satisfactory thesaurus for use in a post-coordinate indexing system, a valuable by-product of the analysis.

4) For single entry systems it is necessary to establish a 'citation order', i.e. the order in which the elements constituting a composite subject are to be stated. Without a fixed citation order it is possible to have 'cross-classification', i.e., the possiblity of placing the same composite subject in more than one location. For example, works on the fractional distillation of mineral oils might be placed in 665.5.048.3 or in 66.048.3:665.5. If the possibility of using UDC as the indexing language in a large international computer-based retrieval system is considered, it will be seen that a standard citation order will be imperative if cross-classification is to be avoided.

Although the need is less obvious, a basic citation order is also necessary in multiple entry systems. Unless the original statement of the subject is semantically sound, manipulation of it to give additional entries is unlikely to be more than partially successful.

A standard citation order already exists in UDC for the common auxiliaries, and further instances will be found in many of the schedules. For example, in Agriculture, Crop is intended to take precedence over Operation or Problem: so that *Insect pest of grain crops* will be found with other works on grain crops, not with other works on insect pests. In Nuclear technology, reactor types are first considered according to neutron energy, then moderator, then coolant, then purpose.

From the generalized statement of facets above we may derive a standard citation order: Things, Parts, Constituents, Properties, Processes, Operations, Agents. It must however be remembered that we may find Constituents of Agents, Parts of Agents, etc, giving a series of recurrences of the basic order. The use of a matrix has been demonstrated by Caless as a solution to this problem.

In many cases, the standard citation order just mentioned will not be appropriate, and it will be necessary to establish one *ad hoc*. The analysis carried out in Stage 2 will prove invaluable here, since in the course of analyzing composite subjects revealed by the literature itself into their elements we are almost certain to gain a clear idea of the relative importance of the different elements. The citation order is merely

a reflection of this importance. The use made of the collection will also be a guide; only those concepts in the facet which is first in the citation order will always be grouped. Concepts in subsidiary facets will be scattered in a single entry system. The choice of citation order is thus very important if the collection is to be arranged to the best advantage.

5) The individual facets can now be arranged to give the overall schedule, since the concepts within each facet have already been arranged. A simple but useful rule is the so-called 'principle of inversion', which states that the order of the facets in the schedule should be the reverse of the citation order. The reason for this is simply to ensure that general topics always precede more special, and it is usually observed in existing UDC schedules. For example, in Agriculture we find that Insect pests (632.7) and Grain crops (633.1) both precede Insect pests of grain crops (633.1-27), so that general is found before special. The filing order puts the Problem facet (including pests) *before* the Crops facet (including grains), while the citation order in the composite subjects where both Crop and Problem are found is Crop-Problem, i.e., the reverse of the filing order.

If this principle is followed, the least important facet of the subject will appear first in the schedule, with the most important last. It is not essential that this principle be adopted (it is not used in the Nuclear reactors schedule, for example), but it does give a more helpful order in that special will follow general consistently; if it is ignored, then special will from time to time be found preceding more general topics. For example, in the nuclear reactors schedule, a work on experimental reactors in general will follow a work on experimental gas-cooled reactors, and both will follow a work on experimental gas-cooled graphite-moderated reactors. This could make searching the files more difficult.

6) It should be possible at this stage to carry out a preliminary test of the schedule to see if it is satisfactory, using the collection of documents on which the analysis is based. This should indicate if the citation order adopted is sound, or if it needs some modification to be satisfactory.

7) The next step is to allocate notation to the schedule. Several points are involved here. The first is to select a base number that is satisfactory and which does not clash with any existing UDC notation. (For private purposes, this base number may be replaced by a letter; as is indicated in the introduction to the English Abridged edition. The possibility of doing this without confusion will give an added check as to whether the subject area covered is indeed homogeneous.)

The second is to allocate the notation to the various facets satisfactorily. At present two symbols are used for auxiliaries within a particular subject: the hyphen- and the point nought .0. It is recommended that the–be used for Parts, Constituents and Properties, and the .0 for Processes, Operations and Agents. An example at the end of this paper illustrates how this may be done in a revised schedule for Library Sciences, in such a way as to give a satisfactory notation for all the more important facets, with the possibility of using the colon if it is thought necessary to make multiple rather than single entries.

The re-use of existing numbers with new meanings is not recommended, since it will lead to possible ambiguities. Existing notation should be left unused, and may then be reused after a period of ten years, by which time the problem will have diminished to manageable proportions in the vast majority of libraries using UDC.

Notation should, as far as possible, be hierarchical or expressive, i.e., each step of division in the schedule should have its counterpart in the notation. This will facilitate the use of the UDC in computer-based systems. UDC gives a decimal notation, i.e., there are nine divisions available of any given number, omitting the zero. In many cases this will not be enough; there will be more than nine divisions of equal standing. In this case, centesimal notation should be used, i.e., two digits (for example 37, 52, 89 etc); this will give 81 potential divisions, or 100 if the zero is not omitted. Since it is desirable to be able to add new concepts, it is necessary not to use numbers immediately together; use 2, 4 and 6 rather than 1, 2 and 3, if only three subdivisions are required.

8) An alphabetical index will be required if topics are to be located easily in the schedule. The simplest way to compile such an index is by what is known as chain procedure; the resulting index is a 'relative index' of the type found in the English Abridged edition. Perhaps the best example is the relative index to the Dewey Decimal Classification, in which the basic principle used by Dewey is worked out very well.

The essential point in this method is that it is not necessary to index subdivisions of a term under that term, since they can be found very simply by turning to that term in the schedules. For example, in Agriculture, Grains are a subdivision of Crops: one would therefore make index entries

Crops: agriculture 633
Grain crops: agriculture 633.1

but NOT

Crops: grain: agriculture 633.1

since the latter merely repeats in alphabetical order the systematic subdivision found in the schedule. The method is very simple, and gives a quick and easy way of compiling a satisfactory index to a schedule. No work is wasted, yet every approach is covered, either by the alphabetical index or by the systematic arrangement of the schedules.

It must be emphasized that what is described here is an index to the schedule, not to any particular collection of documents. It should not be necessary to index any composite subjects, since these should have been excluded from the schedule. While the compilation of an index to a large collection of documents can be complex, the compilation of an index to a schedule is a relatively straightforward matter.

9) Once the index has been compiled, work on the schedule is complete and it should be tested again in detail. It is at this stage that it is useful to circulate the proposal in draft form so that the testing can be shared by other librarians, who may well be in a position to offer helpful comments on the basis of their own experience.

It must be remembered that since knowledge does not stand still, no schedule can ever be regarded as permanent. The analytico-synthetic approach outlined here will give a schedule which may be modified with the minimum of difficulty, since it provides a structure which is likely to endure, even though it may need additional concepts from time to time in order to keep it current.

Outline Schedule Showing an Application of Facet Analysis to Library Science

In order to avoid confusion with the existing notation, this proposal uses 029, at present vacant.

LIBRARY SCIENCE

029	Library science
029.01/.03	Common operations and agents, e.g.,
.01	Controls
	external: legislation
	internal: rules and regulations
	Management
	Finance
.02	Staff
.03	Buildings and equipment (Site, branches, rooms, mobile libraries; lighting, heating, ventilating; special equipment, e.g., shelving) Protection, maintenance, repair.
029.04/.09	Technical operations and agents
.04	Selection, acquisition, etc.
.05	Storing, binding
.06/.07	Classification and cataloguing
.06	Classification
.07	Cataloging
.072	Catalogs by physical form
.074	Catalogs by type of access

.074.2	Author
.074.5	Subject
.074.52/6	By mode of arrangement
.074.52	Alphabetical
.074.54	Classified
.074.56	Alphabetico-Classed
.08	Circulation
029.1/.7	Library services and materials
029.1	Conditions of service, e.g.,
.12	Hours of opening
.14	Reference, lending
029.2/.3	Information-bearing materials
.2	by form (e.g., Books, newspapers, cutting microcopies records, films)
.3	by subject (Use colon, e.g., Law materials 029.3:34)
029.4/.7	Library services by owner, etc.
029.4	by owner
.42	Private: industrial firms, societies, etc.
.44	Academic: universities, schools
.46	Government: municipal, county, state
029.5/.7	Users
or,	
029.4	Special
029.5	Academic
029.61	National
.62/.67	Public
029.7	Special users

Examples of synthetic notation:

Law materials in academic libraries	029.5-3:34
Hours of opening in academic libraries	029.5-12
Classification of law materials in academic libraries	029.5-3[:34].06
Classification in academic libraries	029.5.06

Alternate Outline Schedule for Library Science (recently developed by J. Mills)

029	LIBRARY SCIENCE
.01/.03	Common operations and agents
.01	Organization: administration and management
.012	Administration and Control,
.012.2	External: legislation
.012.4	Internal: rules and regulations
.012.6	Council, Committee, Trustees
.014	Management
.016	Financing
.02	Staff
.022	Professional
.024	Non-professional
.03	Buildings and equipment

.032 Site
.033 Services
.034 Lighting. *By*:628.9
.035 Heating and ventilating. *By*: 697
.036 Fixtures and fittings
.037 Spaces
.04/.09 Technical operations and agents
.04 Accession
.042 Selection
.044 Acquisition
.05 Storage and retrieval
.052 Storing
.054 Binding
.06/.08 Indexing: classification and cataloguing
.06 Classification
.07 Cataloguing
.072 Bibliographical description
.073 Catalogues by physical form
.074 Catalogues by type of access
.075 Author and title
.078 Citation index
.08 Subject
.082 Index and retrieval languages
.083/.085 Pre-coordinate
.083 Alphabetical
.084 Alphabetico-classed
.085 Classified
.085.2 Retrieval languages for
.085.2 (A/Z) Individual systems A/Z
.087 Post-Coordinate
.087.2 Retrieval languages for: thesauri
.087.2 (A/Z) Individual systems A/Z
.088 Automatic indexing
.09 Circulation
–2 Library services and materials
–22 Conditions of service
–222 Hours of opening
–224 Reference
–225 Lending
–227 Open access
–228 Limited access
–24 Ancillary services
–242 Abstracting
–244 Translating
–246 Copying
–3/–8 Service by information-bearing material supplied
–3 By form
–32 Book, monograph
–33 Serial
–34/–36 Non-book

–34	Graphic
–342	Film
–35	Aural
–36	Microform
–37	By provenance
–372	By place of publication.
–374	Publisher
–374.2	Government
–8	By subject. *By*:
029.2/.8	Service by owner-cum-user
.1	Private (individuals)
.2	Special. *By:*
.3	Academic
.44	Primary education
.46	Secondary education
.50	Higher education
.6	National
.7	Local public
.8	Special users. *By*:

Examples of synthetic notation:

Law materials in academic libraries	029.3-8:34
Hours of opening in academic libraries	029.3-222
Classification of law materials in academic libraries	029.3-8:[34].06
Non-Professional Cataloguing staff in academic libraries	029.3.07.024

PILOT SCHEDULES

A POSSIBLE MECHANISM FOR LARGE-SCALE REVISION OF CLASS 55 EARTH SCIENCE) (A.C. Foskett)

It is clear that if one accepts the idea that UDC justifies the expenditure of a considerable effort to make it suitable for use in modern retrieval systems, it is essential to make sure that all the effort expended is directed towards the same end. This will involve a considerable political as well as classificatory effort. A great many librarians have many documents classified by UDC old-style, and will look with no favor on changes which will make their arrangements obsolete. Not all users are convinced of the value of the analytico-synthetic approach; for many of the older generation of users, the straight-forward poppa-momma use of the colon represents the preferred method rather than one adopted for lack of a better. Proposals for change will have to be presented in such a way as to arouse the minimum of hostility, otherwise they stand little chance of being adopted. For this reason, the suggestions put forward here are intended to fit into the present structure of UDC without any significant changes to that structure; they are designed to channel future developments in the direction we would wish to see without alienating present users.

It is also clear that, taking into account present trends, one should envisage a situation where the scheme is used in a very large-scale sophisticated computer-based retrieval system, of international standing. The computer may in fact be the solution to the problem found both by Otlet and LaFontaine and by Bradford (who also started to compile a universal index to recorded knowledge): that, though the problem of acquiring information is difficult, it is minimal compared with the problem of ensuring that the stored

information is used intensively enough to justify the cost of storing it. Both previous attempts were hampered by the physical medium available to them—the card catalogue, which can only exist at one place; the computer provides a means whereby the results of searches are available to anyone, anywhere, in the form of print-out, or even more immediately as a console display.

Many computer-based retrieval systems use 'natural language', i.e., the language of the documents of their abstracts. This involves no intellectual effort at the indexing stage, but research has shown that even in a computer system some control of the vocabulary used improves the results at the searching stage. Some of the systems now functioning with 'natural language', such as that of IBM, may in fact be said to use a controlled vocabulary, since the input consists very largely of documents produced within the organization and therefore written in its input terminology. Furthermore, in any international system, words as such cannot be used because of the language problem; some sort of coding has to be used, so it seems obvious that one should use a coding which may itself be helpful. A classification scheme provides such a coding; of existing classification schemes, UDC seems to offer the greatest potential for development. DC and the Library of Congress classification are both mainly intended as means for the shelf arrangement of books and are not specific enough for the task envisaged. Two other possibilities exist: Colon classification, and the new classification being developed in Britain under the guidance of the Classification Research Group (CRG).

The new CRG general classification will be based on a number of modern theories. Order within specific subjects will be based on principles of facet analysis; overall order will reflect other ideas such as the theory of levels of integration. The possiblities seem very exciting, but a great deal of work remains to be done before the scheme goes beyond the embryo stage, and also it is—at least to begin with—intended for the classification of books rather than smaller units of information. There is, however, no reason why the theories being developed should not be applied in other schemes, such as the UDC.

Colon Classification (CC), the work of Dr. S.R. Ranganathan, is the only general scheme compiled completely on analytico-synthetic principles. Its potential for development has been shown by schedules worked out for very specific areas of technology such as Reciprocating Internal Combustion Engines but at the moment the amount of such development is very limited, and there are large areas of technology for which only outline schedules exist. However, because the scheme is not widely used, large-scale revision along the lines to be proposed here would meet with relatively little opposition, and would almost certainly be welcomed by Ranganathan and his associates. ***Should it prove impractical for political reasons to revise and develop UDC in the desired manner, serious consideration should be given to the possibilities of developing Colon instead.***

As has been suggested, there are political reasons why a proposal for an immediate massive reorganization of UDC would meet with considerable opposition. It is therefore proposed that a more practical plan of campaign would be as follows:

1. The publication of a guide demonstrating to potential revisers the methods they should use.
2. The rationalization of the various notational devices now used for synthesis in UDC.
3. The development of a new set of common subject sub-divisions to replace the existing .00 Point of View numbers.
4. The development of more precise methods of expressing relationships.
5. The construction of a completely revised schedule for the Earth Sciences, which should become the Planetary Sciences, to demonstrate on a fairly large scale the kind of revision needed.
6. The extension to the whole of Science of the ideas developed in 5.
7. The further extension to the whole of knowledge of the ideas demonstrated in 5 and 6.

Some of these steps can be taken now, with a minimum of additional effort; others, notably 6 and 7, are obviously long-time projects involving a great many people, and are beyond the realm of the present proposal. More detailed consideration is now given to those parts of the overall plan which seem to be within the scope of this proposal.

1) Revision work on UDC schedules is going on continuously, yet there is no technical guidance to those involved as to how they should set about compiling a new schedule. There is a document setting out revision procedures, but it is purely administrative and does not give any help in the intellectual effort involved. An outline of the kind of document envisaged has been prepared by Foskett (see page 23) and should be submitted eventually to the CCC. In essence this is an outline of the analytico-synthetic process, and its adoption would at once eliminate many of the problems, at least as far as new schedules are concerned. There will however remain all the existing schedules, many of which do not fit into this pattern and will therefore need to be revised at some future date. Since a completely revised UDC cannot be expected to rise fully-formed like Venus from the waves, the existence of anomalies, some of them serious, will have to be tolerated for some years; such anomalies do not appear to have troubled UDC users in the past, but they may cause some temporary problems in a computer-based system.

Consistent development along the lines indicated would reduce the need for the colon to serve as a jack-of-all trades as it does at present.

2) Freeman and Atherton, in their work for the American Institute of Physics (AIP) found that many of the notational devices used in UDC did not lend themselves to computer searching. For example, the point . is normally used as a separating device, to break up what would otherwise be psychologically unacceptable blocks of figures; if however the point is followed by one or two zeroes (.0 or .00) it does have a significance. The colon can have a variety of meanings, as has already been mentioned. The equals sign = has one meaning used on its own, another, rather different, meaning if it is preceded by a parenthesis (=. For the AIP Project, these signs were replaced by letters to simplify the programming. Such a drastic change is unlikely to win acceptance among the majority of UDC users, and may indeed be unnecessary if the more consistent use of the existing symbols which would result from the application of standard revision procedures become reality. In addition, the adoption of a scheme of relators, discussed under 4 below, might completely solve the problem of the colon.

Two symbols which at present are used without any very great distinction are the hyphen -and point 0. It is not possible to see any rules governing their use from a survey of the schedules. In the Introduction to the English abridged edition, it is pointed out that in Technology the -divisions are used very widely, but this is not true of Science; in the Humanities and Fine arts, it is the .0 divisions which are most widely used. If the suggestions outlined by Foskett (see page 23) are put into effect, the .0 divisions will be those of least significance and therefore of widest application, while the hyphen will be used for Parts, Constituents and Properties, which are likely to be more special to the subject being developed. It is difficult to foresee the exact effects of a more consistent use of these devices, but it should lead to an overall improvement in schedule notation. There may however be occasions when the proposed usage will conflict with established usage; these should be judged on their merits and amended if necessary.

3) The .00 divisions, now called Point of View numbers, are misnamed, and there are many occasions when they conflict with enumeration in particular schedules. For example, .001 is Theoretical Point of View; but theory is frequently listed under particular subjects using other notation, e.g., Psychological Theories 159.9.01; Insurance Theory 368.01; Philosophy of Mathematics 51.01; 521 Theoretical Astronomy; 53.01 Physics Theory; 530.1 General Principles of Physics; 535.1 Theory of Light; etc. Similar conflicts arise with the other topics listed at this point. However, there are also a number of common concepts which do not appear here, e.g., Mobile (Libraries, Grocers' Stores, X-ray Units, etc). What is needed is to change the name, in order to clarify the function, and to expand and redevelop the schedule, at the same time eliminating conflicting pieces of notation from the main schedules. A good model to use would be the common subject subdivisions developed by the British National Bibliography. Subdivisions found in this revised schedule would not normally need to be worked out anew for the revision of any of the main schedules; the only situation where this is likely to be necessary is under equipment, where the general concept 'equipment', will be found in the common subjects facet but will need to be developed in different ways in different subjects. For example, the kinds of equipment used in medicine would not be appropriate to, say, automobile engineering. The development of such differential facets should not prove difficult; all that is necessary is to provide a base in the common subjects facet (as is in fact the case at the moment, where .005 Installation, Equipment Point of View is not developed beyond the general heading).

4) The colon as used now means 'in relation to' but gives no further precision. A system of 'relators' has been developed by Perreault and submitted to the CCC; though this was published earlier this year as an "experimental" P-Note, to be surveyed over over a period of five years, a more recent P-Note seems to indicate its temporary suspension. Perreault's Scheme as it stands conflicts with already existing divisions of Time, e.g., before "711", and potential developments of the Place facet, e.g., above (not at present in Place-1 (orientation), but obviously belonging there), as well as concepts such as "outside" (-194) which are already in the Place facet in UDC. There is a need for a detailed study of relators; not merely the Perreault Schema, but also the modification of it proposed by Wesseling and the effect that any such scheme would have on the existing schedules. It may be that a more satisfactory approach to the development of the main schedules would eliminate at least some of the problems which the Perreault Schema as it now stands tries to solve; in a sense, all concepts in a classification scheme stand in some sort of relationship to other concepts, and as yet there does not exist any theoretical basis which would make possible a clear distinction between those which can be built into facet relationships and those which need to be developed in isolation. Farradane's work may be relevant here.

5) The foregoing may in a sense be regarded as a prolegomena to the core of the proposed Project, which is to demonstrate that a completely new schedule for a fairly large basic area in Science can be drawn up in such a way as to give results which are sound not only from the point of view of the subject but also from the point of view of classification theory. It has already been pointed out that perhaps the major opposition that will be encountered will not be on the grounds of classification theory but of departure from precedent: that librarians and other users will be unwilling to accept major changes if these mean large-scale reclassification. A complete revision of Science is unlikely to be approved, and to attempt it would probably lead merely to frustration. On the other hand, the Earth Sciences lend themselves well to the design, in more than one way. Dewey allocated one 'division' of his notation to Geology: 550-559; however, he only used part of this to develop schedules for the subject, leaving 554-559 for division by Place. UDC had no need of these divisions, as there is a distinct Place facet which can be used as required; although 556 has recently been developed for Hydrology (incidentally taking this subject out of its context in Geomorphology), 559 is still vacant, and could be used as the base number for a new schedule.

Man has now reached the point where the name "Earth" sciences is something of a misnomer, and the opportunity could well be taken to develop the new schedule under the title Planetary sciences. This would have additional consequences of some value. The overall existing order for Science in UDC is as follows:

Science
- Mathematics
- Astronomy
- Geodesy. Surveying
- Physics
- Chemistry
- Earth Sciences
- Palaeontology
- Anthropological and Biological Sciences

Whatever theoretical approach one takes, it is impossible to justify the placing of Geodesy, and in fact Astronomy is also out of place if one follows a progression of dependence – it should follow both Physics and Chemistry. It can also be argued that Planetary Sciences should follow the Biological Sciences, though the progression here is less clear. Be that as it may, to develop a new schedule at 559 for Planetary Sciences would make it possible to remove the anomalies from the Astronomy schedule, incorporating them where they belong. The use of 559 would enable the new schedule to be published as an "experimental" P-Note, if this was thought desirable, though in fact the proposed schedule is unlikely to be so startlingly different from other UDC schedules of recent date as to require the designation experimental.

The development of this new schedule would free the whole of the rest of 55, for redevelopment after a starvation period. One would envisage Astronomy being redeveloped at 558, Biological Sciences at

555, 556 and 557, Mathematics at 551 and Physical Sciences at 552, 553 and 554. *After the required starvation period, all that would be necessary would be to drop the initial 5, and the result would be a complete redevelopment of Science.* This of course a long-term project; one would need to think in terms of over ten years for the final redevelopment, by which time the revision may have taken a different direction. Nevertheless, it represents a means by which such a revision could be carried through with a minimum of disruption to existing services using UDC.

The new schedule would probably include the greater part of what is not included at 55, the only exceptions being those individual concepts now included but belonging elsewhere, as for example Earthquake Disasters was excluded in Seismology. It would also include Mineralogy, now at 549, some parts of Geodesy, now in 528, and some parts of Astronomy. The work would require a collaborative effort involving both subject experts and classificationists, to devise an outline schedule, using the principles of facet analysis or the ideas about entities and attributes now being developed by the British CRG. This could then be studied by subject experts, from the point of view of overall order, relation to the scientific and educational consensus, and comprehensiveness. Once agreement had been reached, each topic within the overall schedule would be developed, following the same procedure. Past experience suggests that the classificationists are more likely to be able to understand what the subject experts have in mind than vice versa; for this reason it would be necessary to go into some detail in explaining just what the schedule is intended to do, and how it is meant to work. Once these principles have been grasped it is usually not too difficult to reach agreement. Some parts of the new schedule have already been prepared, for example Seismology; here, all that will be required is some changes in the notation, but this is trivial.

6) It has been shown above that the development of a new schedule for Planetary Sciences could be the starting point of a much more massive revision. While it does not seem practical to start work on this at this point in time, it will be important to see that no other revisions are accepted that could conflict at some future date with the kind of reorganization envisaged here.

7) A massive reorganization of the *whole* scheme is likely to take a great many years. The practical steps that can be taken now relate mainly to the rationalization of the use of the notational devices, the development of a system of relators, and the psychological preparation of users for the change. The amount of effort that will be involved should not be underestimated; UDC is a large scheme, with a large number of users, many of whom are not sympathetic to the ideas underlying this proposal. The need for the development of international information services is surely great enough to warrant the expenditure of such an effort.

DEVELOPMENT OF PLANETARY SCIENCES CLASS IN COLON (D. Langridge)

The Earth Sciences are included in Class H of Colon Classification; called Geology. It includes Mineralogy, Petrology, Structural geology, Dynamic geology, Stratigraphy, Palaeontology, and Economic geology. The detail is only sufficient for classifying books, so that a full scale enumeration of isolates would be called for in making a scheme suitable for documentation. If the present structure of the class is unsatisfactory it could be changed. It might also prove necessary to add some detail to related classes such as Chemistry.

Whatever proved necessary would be possible, since Colon theory now provides the material to accommodate any subject, however detailed. Any research worker undertaking such a task would first have to absorb thoroughly the principles of the scheme. This would be most quickly and efficiently achieved by instruction from someone with experience. It would be unwise to attempt it without such help.

He should establish contact with those in India who have been doing similar work on other subjects.

History

The idea of Colon Classification was conceived in 1925, when Dr. Ranganathan studied librarianship in England. He saw immediately the limitations of the Dewey scheme in dealing with compound subjects

and realized that the solution was complete analysis of every class into categories of terms. He drafted the new scheme before he left England and during the next few years developed it in application to the collection in the University of Madras library. The first edition was published in 1933.

The scheme was adopted by other Indian libraries and is now used by the universities of ten Indian states, by nearly all public libraries in Madras and Maharashtra states, by some public libraries in other states, and by some government departments. It has also been used for classifying published indexes to scientific periodicals, and the bibliography by Das Gupta[6].

Unlike all the other general schemes, Colon has not been conservative in its revision policy. Ranganathan has made quite radical changes in successive editions to improve the scheme and to meet the changing demands of published knowledge. The latest edition is the sixth, published in 1960, with amendments in 1963. The seventh edition is in preparation.

This published version of Colon is known as the basic classification and is intended for organising books and catalogues in general and special libraries. It does not aim to give the specificity required for documentation work, and should therefore be compared with the abridged, rather than the full edition of UDC.

Since the mid 'fifties, Ranganathan has concentrated on the problems of depth classification. The structure and notation of Colon have been developed to deal with the most detailed specification that any documentation service would require. The methodology was described in the first issue of *Library Science*. . . .(March, 1964). Detailed schedules for Agriculture and Management had already appeared in issues of *Annals of Library Science*. Since 1964, *Library Science*. . .has included schedules for such subjects as Nut production engineering, Medical radiology, Locomotive production engineering and Glass production technology.

Work continues in India under Ranganathan's guidance. Progress is inevitably slow since there are few research workers engaged and there are many subjects in science and technology requiring detailed schedules.

Although the Colon scheme itself has not been adopted by libraries outside India, it has had considerable influence on theory and practice elsewhere, particularly in Great Britain. The Classification Research Group began in 1952 with Ranganathan's work as their main influence although not committed to Colon Classification itself. Special schemes made by CRG members and others since that time are closely related to Ranganathan's fundamental principles. The UDC has incorporated some ideas from Colon and revisions of individual classes have benefited from the strict application of categories and citation order (first demonstrated in Colon). Similar methods have been applied to the maintenance of the Bibliographic Classification in the *BC Bulletin*, and at the British National Bibliography classifying by Dewey is optimized by the application of Colon discipline. The new general classification being developed by the Classification Research Group, with the primary aim of computer retrieval, is based on virtually the same fundamental categories as Colon.

In fifteen years of research and teaching in classification, I have found Colon by far the most fruitful source of theory and exemplification of practice. Considerable use of the scheme in teaching has also convinced me that once the principles are mastered it is the easiest of the general schemes to use and the most efficient. The general failure to recognize this in the West, I attribute to lack of experience in applying the scheme. The common criticism of Colon's notation is almost certainly exaggerated, and in any case does not apply to *computer* retrieval.

Main Features of Colon Classification

1) It should first be emphasized that there is nothing specifically Indian about any of the *essential* characteristics of Colon. The scheme is as suitable for use in the West as it is in India. Any shortcomings in provision for particular subjects, or terminology not current in the West, are not fundamental characteristics and can easily be modified.

2) The most important feature of Colon, and the one that gives it a marked superiority over all other existing general schemes, is the fact that it is based on a complete and coherent body of theory which has been steadily developed by Ranganathan side by side with its constant application and testing in practice. This theory is fully expounded in the "Prolegomena."[7]

3) Colon is the only general scheme that is completely analytico-synthetic. UDC suffers from being developed piecemeal from an enumeration scheme (as well as from the lack of a complete and coherent body of principles). In my recent consulting work for the George Washington University, it has been very obvious that all the difficulties in adapting UDC schedules for a computer are due to this lack of a complete system. Furthermore, the solution to these problems can always be found in Colon theory and practice.

4) Colon theory is based on a careful distinction between the different problems arising in the three planes of work—Idea plane, Verbal plane and Notational plane.

In the Idea plane the significant elements are Basic classes, Categories, Rounds, Levels and Isolates.

Basic subject. These are the areas of interest into which the whole of knowledge is first divided. The principle is the same as that used in the UDC.

Isolates. These are the elementary terms referring to observed phenomena—whether objects, activities, properties, etc. They do not exist as subjects until combined with a Basic subject; e.g. the isolate Gold is associated with Basic classes like Chemistry, Mining, Economics, etc. to give concrete subjects. Common isolates are those applicable to all or many Basic classes, e.g. Evaluation, Institution, Bibliography. Special isolates are peculiar to one Basic class.

Categories. Each isolate is assigned to one of five fundamental categories which provide a common pattern for the whole Colon scheme. The categories are named Personality, Matter, Energy, Space and Time. More precise terms have been used by other classification theorists working in limited areas of knowledge. There need be no incompatibility between different sets of categories. The advantage of Ranganathan's set is that they are the most generalized, and therefore the most widely applicable. Once their meaning has been grasped by studying Ranganathan's theory and practice they are not difficult to use. It is interesting to note that the Classification Research Group, who refused to accept Ranganathan's categories on trust, have independently arrived at three main categories for their research on a general scheme which correspond very closely to Personality, Matter and Energy. The CRG terms are Entities, Properties and Activities. This was after considerable experience in using more specialized categories for the construction of schemes for a wide variety of subjects.

The citation order of categories for any subject is always the same in Colon, viz. PMEST.

Rounds. The theory of rounds is a necessary complement to the categories. In any compound subject, the same fundamental category may occur more than once, e.g. in Medicine the first round (P) is Part of body; second round (P) is agent of disease; third round (P) is agent of treatment. Another round is always initiated by the Energy facet, thus we may have PME 2P 2M 2E 3P 3M 3E . . . S T. Space and time always occur once only, in the last round.

Even the more detailed categories recommended by Vickery[8] for use in science and technology do not account as clearly for this recurrence of categories. Ranganathan's theory is essential.

Levels. The theory of levels provides mainly for the Whole/Part relationship. It states that within any given round there may be more than one level of a category, e.g. in Botany the first level is whole plant, the second level is Parts, such as leaf or root. This applies to all categories except Energy. Levels are symbolized as P1, P2, P3, etc.

5) *Notation*. The adverse comment on Colon notation has been related solely to human searching; but those very factors that make the notation complicated for people also make it eminently suitable for machines.

Categories in Colon are introduced by distinctive symbols, thus showing clearly the facet structure of individual subjects. Within facets the hierarchy of genus/species relations is shown as far as possible by the notation. Since it is impossible to preserve a pure hierarchical notation, some conventions are necessary. Ranganathan provides these largely in the form of Sector notation of which octave notation (used also in UDC) was an early example; e.g. in any class x, x1-8, 91-98, 911. . . symbolize coordinate classes.

Ranganathan has a bigger range of devices for interpolating new subjects than any other scheme. His notation reflects more precisely the structure of the subject than that of any other scheme and would therefore pose the least problems in computer programming.

SEARCH STRATEGY

PARENTHESIS-FREE NOTATION FOR COMPUTER SEARCHING THE UDC (T.W. Caless)

If we designate THING/KINDS as Main Facets (M) and PARTS through AGENTS (plus PLACE, TIME and FORM) as Subsidiary Facets (S), then for ease in handling, we can symbolically show conventional UDC numbers (or even natural language indexing strings) assigned to documents as combinations of M's and S's as: (each developed in the established facet order)

Document #1–$M_1S_1S_2S_3$
Document #2–$M_1S_1S_2$
Document #3–$M_1S_1S_2S_3S_4$
Document #4–$[M_1S_1S_2] + [M_2S_3S_4]$
Document #5–$[\,[M_1S_1] + [M_2S_2S_3]\,]\,S_4S_5$

where the subscripts show only the difference between first appearing terms, second appearing terms, etc., in the indexing strings and the square brackets are used as purely an algebraic device. It can be seen that Documents #1 thru #3 represent reasonable precise specification of the subjects by having 3-5 components or terms. Each one of these terms can itself be a file approach (access) term. Until now, the only foolproof procedure for this was to permute each term exhaustively. This is, however, quite expensive, especially for cases where the analyses have been taken to the indexing depth shown in Documents #4 and #5. Some investigators have avoided exhaustive permutation by using chain indexing procedures which has resulted in a poorer performing retrieval capability.

This paper addresses the technical problems of demonstrating that it is both technically and economically feasible to computerize the UDC to any depth of indexing that is required. The multi-dimensional concepts found in fully intact UDC numbers can be retained for both filing and searching purposes by translating the UDC notation into parenthesis-free notation (known as Polish Notation[9]). An additional advantage of this method is the ability to keyboard intact UDC numbers directly into the computer file; the programming will automatically translate the UDC numbers into Polish Notation.

It has been found that UDC numbers and natural language indexing strings, when translated into Polish Notation, can be effectively searched intact with every term in the indexing string serving as an approach term and with no loss of concepts displayed in the string. Polish Notation has been used extensively for handling complicated notation on computer files[10], and a Polish expression is essentially an Operator, a left Operand and a right Operand. Each Operand can additionally be a Polish Expression which gives it the power to handle any complication found in a UDC number. For example, if we use the Polish Expression as our fundamental analysis element, and let o, the Operator, be any facet indicator, and A, the left Operand be any indexing component or term and B, the right Operand, be any different indexing component or term, we have:

Referring now to Document #1, we can translate

$M_1S_1S_2S_3$

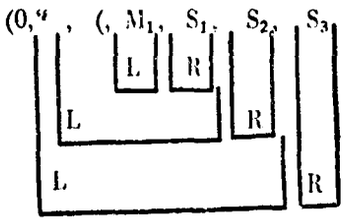
into Polish Notation only if we know what the operators (facet indicators) are between indexing components and what combining procedence (or order) they have. If we assume that during the subject analysis we assigned M_1 from some category in the UDC Main Classes and selected S_1, S_2, and S_3 from the Place, Time and Form Common Auxiliaries, we would know from Perreault (see pages) that their respective combining precedence would be 4, 4, and 2, and:

Example 1

(4) (4) (2)

$M_1(S_1)$"S_2"(OS_3)

where the underline grouping indicates the $M_1(S_1)$ combines first, "S_2" combines with it next and (OS_3) combines last. Converting this string into Polish Notation produces:



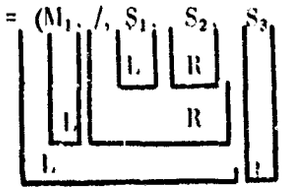
Note that commas are used here to separate the elements, the R's and L's signify right and left Operands respectively and the closing parenthesis and quotation marks have no role in "Polish" and are eliminated. Conversely, if our subject analysis revealed that our S_1 and S_2 were Common Auxiliaries of Place consecutively extended over a range of more than one location and S_3 was the Language Auxiliary, we would have:

Example 2

(4) (8) (1)

$M_1 (S_1 / S_2) = S_3$

where the combining precedences are considerably different from the first example. Translating this into Polish Notation, we now have:



Examining Example 1 and 2 closely will quickly reveal a significant difference in the two indexing strings and some indication of how complicated UDC numbers can be. But by translating the UDC numbers into Polish Notation for filing and search purposes, we essentially linearize the complications and make it possible to computer search effectively under any conditions. Using these two examples, we will assume that we have been requested to find any document that contains an M and an (S). And in Example 1, the M_1 and the (S_1) surrogates match precisely, signifying that this document should be a "hit." The search request is constructed (in an order determined by our facet order) and shown in a symbolic UDC number as:

M(S),

and translated into Polish Notation as:

so the search can proceed. The computer sequentially looks for an opening parenthesis which it finds after rejecting the (0, and "-". This partially satisfies the search request. Proceeding further into the indexing string, the computer looks for an M which is a left Operand, finds it, and proceeds to look for the S, a right Operand, which it finds and the document is retrieved. In Example 2, the M exists in the indexing string but we will assume that the (S) in the search request corresponds to the surrogates in S_2 and not S_1. We can judge by inspection that the computer should retrieve this document. The search begins by the computer rejecting the - but accepting the (. Looking further, the M is matched and is identified as a left Operand which further partially satisfies the search. The computer continues looking for a match for S, rejects the / and S_1 but accepts the S_2, a right Operand, which completes the search and the document is retrieved. It is of interest to note that in Example 2, if the S_1 surrogate matched the search request instead of the S_2, it too would have satisfied the computer search since even though it is a left Operand of the Polish Expression /S_1, S_2, it is a member of a right Operand in the Polish Expression (, M_1, / S_1, S_2.

Obviously, even though the M's and S's shown above were demonstrated with UDC surrogates, the Polish Notation could additionally handle natural language indexing strings, regardless of their length. The only justification for doing this would be if the indexing strings contained multi-dimensional concepts, as do the UDC strings. There would be no need for this sophistication, however, if the indexing was unregulated and the analysis did not provide categories (facets) for the subjects to be handled.

[1] Farradane, J., et al., 1966. "*Research on Information Retrieval by Relational Indexing*," Part I, METHODOLOGY, The City University, London. 60 pp.

[2] Perreault, J.M., 1965. "*Categories and Relators: A New Schema*," Rev. Int. Doc. 32(4):136-144.

[3] Caless, T.W., 1969. "*Subject Analysis Matrices for Document Classification*", Classif. Soc. Bul., 2(1):29-37.

[4] Caless, T.W., 1969. "*Subject Analysis Mat for Classification with the UDC*", Proc. of First Seminar on UDC in a Mechanized Retrieval System FID/CR R(t No. 9, Copenhagen. 4 p.

[5] Austin, D., 1969. "*Development of a New General Classification: A Progress Report*," The Information Scientist, November. pp 95-115.

[6] Das Gupta, A.K., 1967. "*An essay in personal bibliography*." Asia Publishing House, (Ranganathan Festschrift Vol. II) includes details of several hundred items relating to Colon up to 1961.

[7] Ranganathan, S.R., 1967. "*Prolegomena to library classification*." 3rd ed. Asia Publishing House.

[8] Vickery, B.C., 1959. "*Classification and Indexing in Science*," Butterworth, London.

[9] Lukasiewicz, Jan, 1961. "*Aristotle Syllogistic from the Standpoint of Modern Formal Logic*," Clarendon Press, Oxford, England. p. 78.

[10] Iverson, K.E., 1962. "*A Programming Language*," Wiley. 286 pp.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
[illegible] Applications Inc. for Jet
The George Washington University
[illegible], D. C. 20[illegible]

2a. REPORT SECURITY CLASSIFICATION
[illegible]

2b. GROUP

3. REPORT TITLE
[illegible] FOR MANIPULATING UNIVERSAL DECIMAL CLASSIFICATION [illegible] FOR COMPUTERIZATION.

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Final

5. AUTHOR(S) (First name, middle initial, last name)
T. L. Caless J. Mills
A. C. Puckett J. M. Perreault
D. Lowrie

6. REPORT DATE
December 1970

7a. TOTAL NO. OF PAGES
40

7b. NO. OF REFS
10

8a. CONTRACT OR GRANT NO
[illegible]-68-C-0035

b. PROJECT NO 9769

9110[illegible]

[illegible]

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AFOSR 70-[illegible]

10. DISTRIBUTION STATEMENT

1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES
TECH OTHER

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research (NM)
1400 Wilson Blvd
Arlington, Virginia 22209

13. ABSTRACT

Summarizes research findings of Principal Investigator and part-time Consultants who conducted an investigation of the Universal Decimal Classification (UDC) as an indexing language for computer retrieval. Problem areas were identified (structure, schedules, application), new developments reviewed (combining [illegible] devices, subject analysis matrices, guidelines for schedule revision, pilot schedules, parenthesis-free notation forsearching), and a broad overview and history of the UDC are presented in ten papers. Faceted classification techniques applied to the UDC are recommended throughout the investigation for better consistency in application.

DD FORM 1 NOV 1473

Security Classification