*Patent Office*
*Research and Development*
*Reports . . . . . No. 12*



# STORAGE AND RETRIEVAL OF CONTENTS OF TECHNICAL LITERATURE

## NONCHEMICAL INFORMATION

*Second Supplementary Report*

*Prepared by*

**Simon M. Newman**
*Staff Member*

**Office of Research and Development**
**Patent Office**

**November 17, 1958**



**Lewis L. Strauss**
*Secretary of Commerce*

**Robert C. Watson**
*Commissioner of Patents*

For sale by the U. S. Department of Commerce, Washington 25, D C. Price 25 cents.

## Table of Contents

## Foreword

This paper constitutes a third interim report on continuing Patent Office research in Ruly English. A lone researcher in close association with others doing related work invariably draws both consciously and unconsciously from them for his ideas and conclusions. Hence the author gratefully and freely acknowledges the contributions of those of his staff associates and others in this and related fields who find their ideas included in this report.

The responsibility for this report, however, remains solely that of the author. He wishes to thank all who have taken the trouble to comment upon previous efforts and welcomes reactions germane to these latest observations.

*Simon M. Newman*

# STORAGE AND RETRIEVAL OF CONTENTS OF TECHNICAL LITERATURE

## NONCHEMICAL INFORMATION

## Second Supplementary Report

### INTRODUCTION[1]

One of the approaches taken by our Office in its research on storage and retrieval of information in technical literature is linguistic in nature, since most of this literature is already verbally expressed, and since drawings, tables of figures, photographs, models, working machinery, etc., can be transliterated into their verbal counterparts.

Any two technical documents which refer to the same or to closely similar phenomena will ordinarily express these phenomena in two widely different sets of verbal expressions, both of which accurately convey the same information to the human mind. In using a mechanized searching system, a patent examiner might formulate his request for the retrieval of this same information by utilizing a still different, but accurate verbal expression. If his search is to yield both documents, he must use some metalinguistic common denominator.

We have therefore undertaken the creation of a metalanguage in which each unit will refer to one, and only one concept, and in which each concept may be expressed by one, and only one unit. We have named this metalanguage Ruly English after the terminology of Professor Dodd,[2] who pointed out that English is quite "unruly." By the use of Ruly English it will be possible to convert the many complex and interrelated notions in technical documents into single unique forms. In its concern for linguistic interrelationships in the retrieval of detailed information, such a scheme as this is believed to be fundamental to the systems now under development in the Patent Office.[3]

The basic elements of Ruly English are itemization, distribution, interfixing, modulation, qualification, and quantification.

Itemization[4] consists of assembling all descriptors of a single thing, each of these descriptors describing the thing from a different aspect, and grouping them as one item in a numbered list. Such numbers are used only for identification.

Distribution[5] is applicable to notions of interaction or interrelation between two or more items. In our searching at the Patent Office we are much more interested in the interaction between and in the interrelation of things than we are in the specific details of the individual things themselves. Such notions are expressed by means of two or more cognate descriptors, one of which is placed with each related item. When necessary, these cognate words may be polarized to designate the direction of interaction or interrelation. Thus distributed, the notion expresses the relation between the items in which the cognate descriptors appear. The cognate descriptors which express interrelation are known as interrelational concepts.[6]

Interfixing[7] constitutes a method of recording the existing network of interactions and interrelations between descriptors. Any two or more descriptors that interact or are interrelated and every set of distributed cognate descriptors are affixed with the same arbitrary, unique number (known as an interfix) signifying their interrelationship.

Modulation[8] constitutes a means to limit the size of the encoding vocabulary. The elemental Ruly English terms are reduced to Roots, and their meanings are modified by adding standardized suffixes called modulants. Hence one Root can be modulated to serve the many different aspects of its basic concept.

Qualification[9] is accomplished by adding a qualifying descriptor to the desired item. In many situations, however, an otherwise normal Ruly descriptor is not satisfactory. For example, many qualifying concepts are located somewhere along a scale between two limiting qualifying terms, but these concepts never fulfill the definitions of the terms at either end of the scale. Qualifying Roots of this type--called dual-aspect--are specially coined by combining the limiting terms into one Root. For example, the combined terms "resilient" and "rigid" become RESILRIG.

Quantification[10] of qualifying Roots is done by means of prefixes, e.g., SLI' (for slightly) in the expression "SLI'MORE," meaning slightly more. When applied to dual-aspect Roots, the prefix quantifies the first one of the two combined qualifying concepts; e.g., "SLI'RESILRIG" is used for "slightly resilient" (thus, "materially rigid").

Before any meaningful retrieval tests involving Ruly English can be made, however, the created Roots, modulants, and interrelational concepts, together with any additional terms necessary for expressing notions in the field used for such testing, must first be critically defined. Secondly, those Roots which include, or are subsumed from, other Roots must be recorded in a series of inclusive-subsumed categories.[11]

Two appendices are accordingly provided to this paper: Appendix A, which includes, inter alia, an alphabetical list of Roots, modulants, and interrelational concepts with definitions; and Appendix B, which consists of an alphabetical list of subsumed-inclusive relations. In addition, Appendix A includes common terms, each paired with the Root or Roots that have been so defined that they in-

clude or are related to such terms. These terms will be interspersed with the Roots and their definitions, and Appendix A will also serve as an alphabetical index for finding an appropriate Root from an "unruly" concept. These two appendices constitute simple prototypes of two of the tools necessary to a rudimentary Ruly English adequate for Patent Office information retrieval.

## DEFINING ROOTS AND MODULANTS

In collecting the Roots defined in Appendix A, previously issued reports on Ruly English[1,2] were scanned for defined terms. Some of the definitions were necessarily revised in the light of other Roots that had been created later. In a few cases, some terms were not included since correct definition, or position in inclusive-subsumed categories, was ambiguous or inaccurate and no clear definition or category was discovered. This process of revision of the "meaning" of notions will continue as long as new Roots for notions are created. Wherever possible, existing Roots, modulants, and interrelated concepts have been utilized as defining terms. Additionally, certain Roots have been adopted without any attempt at definition, because it is believed that the notions conveyed by these terms are well understood. At some future date, when sufficient Roots have been collected and defined, it should be possible to make a statistical study of the non-Ruly words used in the definitions and to derive from them a limited set of undefined terms by which all future notions could be defined.

## SUBSUMED—INCLUSIVE RELATIONS

The alphabetic list of Roots with their more inclusive terms, collected in Appendix B constitutes a first approximation of one means for illustrating the subsumed-inclusive relations of Roots. Although the Patent Office has historically indented the species under the genus, this positional arrangement is reversed in this appendix since the main use of the list will be to find more inclusive terms. An example showing how this appendix was set up will, I hope, prove helpful.

To a zoologist, cats and dogs are categorized as mammals, while to some owners they are categorized as pets along with guppies and canaries. To other owners, a dog might be categorized as a form of burglar-protection along with window-guards and door-locks, while a cat could be categorized as a rodent-destroyer along with mouse-traps and arsenic. To the owners' neighbors both cats and dogs might be categorized as nuisances along with trucks and salesmen. For more inclusive terms, I have taken living-things, which subsumes both mammals and pets; and mechanisms which subsumes both trucks and door-locks. If all of these terms were Ruly, a list for the Appendix would be set up as shown in Schedule 29.

Here we see that dog is subsumed under burglar-protection, mammal, nuisance, and pet, these latter terms being indented under dog; both mammal and pet being each further subsumed under living-thing; and, additionally, in alphabetic order, we see both mammal and pet repeated with its more inclusive and indented term living-thing.

SUBSUMED-INCLUSIVE RELATIONS

Arsenic
  Rodent-destroyer

Bird
  Living-thing

Canary
  Bird
    Living-thing
  Pet
    Living-thing

Cat
  Mammal
    Living-thing
  Nuisance
  Pet
    Living-thing
  Rodent-destroyer

Dog
  Burglar-protection
  Mammal
    Living-thing
  Nuisance
  Pet
    Living-thing

Door-lock
  Burglar-protection
  Mechanism

Fish
  Living-thing

Guppy
  Fish
    Living-thing
  Pet
    Living-thing

Mammal
  Living-thing

Mousetrap
  Mechanism
  Rodent-destroyer

Pet
  Living-thing

Salesman
  Mammal
    Living-thing
  Nuisance

Truck
  Mechanism
  Nuisance

Window-guard
  Burglar-protection
  Mechanism

Schedule 29

The mere mechanics of setting up a complete list such as this is staggering, but mechanization of this process is undoubtedly possible. In practical use, it also may be possible to restrict the list to one indentation, since every subinclusive Root which has one or more Roots indented thereunder is repeated as an original entry. If this shortcut were used, however, the list would not reveal the complete hierarchy of any one Root. Since such hierarchies are very helpful in defining other Roots, at least for the present the full hierarchical list of indentations for each inclusive word will be listed.

## PROCESS, APPARATUS, AND WORK RECONSIDERED

By combining several definitions from Appendix A, the PROCESS modulant, =NT, becomes: an action or behavior which characterizes either (1) a property, existing circumstance, activity or mode of being, or (2) a change in the property, existing circumstance, activity or mode of being, of one or

more SUBSTANCEs, things, or intangibles. The change of property, existing circumstance, activity or mode of being is known as a change in STATUS,[13] and most MANUFACTURing[14] PROCESSes are concerned with such STATUS changes.

Upon analysis, most of these changes of STATUS usually occur concurrently. One species of such concurrent PROCESSes involving dual-aspect Roots (e.g., HEATCOOL) was referred to previously.[15] Further analysis of these concurrent PROCESSes reveals that two interacting objects are usually involved.

MANUFACTURing has many other typical concurrent PROCESSes. At the time that a metallic bar is cut with a shear, a number of concurrent PROCESSes occur, e.g.: (1) the shear PARTs the bar, (2) the bar dulls (RESHAPes) the shear, and (3) some of the energy used in moving the shear is transduced as the speed of the dies is reduced to raise the temperature of both the bar and the shear. Setting aside for a moment the third or energy transduction PROCESS, the action might be itemized as shown in Schedule 30, wherein both the bar and the shear have descriptors of both WORK and apparatus. The interfix 1 shows that the bar constitutes the WORK and the shear constitutes the apparatus in the PARTing PROCESS, while the interfix 2 shows that the shear constitutes the WORK and the bar constitutes the apparatus in the RESHAPing PROCESS.

| Item No. | Root, Modulant, and Interfix |
|---|---|
| 1 | Bar<br>PART=W-1<br>RESHAP=OR-2 |
| 2 | shear<br>PART=OR-1<br>RESHAP=W-2 |

Schedule 30

Another concurrent PROCESS is exemplified in the coining (RESHAPing) of a single piece of metal by impacting it between two dies. This situation was deliberately chosen because it has the following characteristics which make it troublesome: (1) It constitutes a non-repetitive PROCESS in which only a minute change is effected upon the metal piece, and (2) the impacting not only RESHAPes the metal piece, but also minutely but permanently deforms (RESHAPes) the dies. Here any distinction between WORK and apparatus becomes almost nonexistent, except as one considers the reason for which the PROCESS is performed, since only one piece of metal is involved and the RESHAPing is so slight as to approach the amount of change in the dies. However a search request might be made for either RESHAPing process, and either can be retrieved if itemized as shown in Schedule 31.

| Item No. | Root, Modulant, and Interfix |
|---|---|
| 5 | piece<br>RESHAP=OR-3<br>RESHAP=W-4 |
| 6 | dies<br>RESHAP=W-3<br>RESHAP=OR-4 |

Schedule 31

It is now clear that apparatus must be classified as a species of WORK. The modulant of apparatus must therefore be =WOR, rather than =OR, and it accordingly should be subsumed under WORK.[16]

It remains now to illustrate how the process of heating the dies can be shown by itemization, as in Schedule 32. Here items 7 and 8 show the bar and dies before the process of RESHAPing has taken place, while items 9 and 10 show them afterwards. The STATUS modulants show the temperatures and speed before (=EP) and after (=ER).

| Item No. | Root, Modulant, and Interfix | Interrelational Concept and Interfix |
|---|---|---|
| 7 | Piece | |
| | RESHAP=NT-5 | TIMAFOR-8 |
| | RESHAP=WOR-6 | |
| | RESHAP=W-7 | |
| | Temperature=EP | LESS-9 |
| 8 | Dies | |
| | RESHAP=NT-5 | TIMAFOR-8 |
| | RESHAP=W-6 | |
| | RESHAP=WOR-7 | |
| | Temperature=EP | LESS-9 |
| | Speed=EP | MORE-10 |
| 9 | Piece | |
| | RESHAP=NT-5 | AFTIM-8 |
| | RESHAP=WOR-6 | |
| | RESHAP=W-7 | |
| | Temperature=ER | MORE-9 |
| 10 | Dies | |
| | RESHAP=NT-5 | AFTIM-8 |
| | RESHAP=W-6 | |
| | RESHAP=WOR-7 | |
| | Temperature=ER | MORE-9 |
| | Speed=ER | LESS-10 |

Schedule 32

## APPENDIX A

### Roots, Modulants, Interrelational Concepts, and Index of Terms

Schedule 33 constitutes an alphabetical list of modulating terms with their modulant symbols.

An asterisk (*) following a Root indicates that the Root is not defined.

RULY MODULANTS (Modulating suffixes)

(Schedule 23, Report #4, Revised and Alphabetized)

| Explanation | Modulant |
|---|---|
| Disease of | =IS |
| Made from (made out of) | =M |
| Combination Including | =MCI |
| Component of Complex | =MCC |
| Source SUBSTANCE | =MSM |
| Numerical | |
| Exactly | =Y |
| Or LESS | =Z |
| Or MORE | =X |
| Used as an Ordinal | =B |
| PROCESS | =NT |
| STATUS, a condition of — ing | =E |
| After PROCESS | =ER |
| Before PROCESS | =EP |
| DURING PROCESS | =ED |
| Subcombination of Whole | =SW |
| WORK | =W |
| Apparatus or Performer | =WOR |
| Composition Descriptor | =WCD |
| Final Product | =WFP |
| Ingredient Descriptor (restricted to chemical disclosure) | =WID |
| Starting Material | =WSM |
| Intermediate Product | =WIP |

Schedule 33

About APPROXTIM( )

Action PROCESS

Adorn ADORNBLEM

ADORNBLEM The pleasing or displeasing visual appearance of a MODULE or ENTITY.

AFORLAP(LAPAFT) A 1=B AFORLAP CONDITION or STATUS begins TIMAFOR a 2=B LAPAFT CONDITION or STATUS and ends AFTIM the start and TIMAFOR the end of the 2=B CONDITION or STATUS, e.g., starting a walk before a storm and ending before the storm is over.

After AFORLAP(LAPAFT)

TIMAFOR(AFTIM)

AFTIM TIMAFOR(AFTIM)

ALIGN The relation of the AXISs of 2=X CONFORMs, or PORTIONs or ELEMENTs of the same or of different CONFORMs.

Alignment ALIGN

Almost time TIMAFOR(AFTIM)

AMONG( ) 3=X CONFORMs are each AMONG the others if, together, they constitute a group.

ANGLE* (geometric term)

cf. RTANG

APPROX( ) Very closely the same value, e.g., the diameter of 2=Y INTERSECTing CONFORMs consisting of a RTCIRCYL in a RTCIRCYLical hole in a force fit.

Approximately APPROX

APPROXTIM( )

APPROXTIM( ) 2=X CONDITIONs or STATUSs are each APPROXTIM when they occur sequentially, but the order of the sequence is not disclosed, e.g., he died about 6 o'clock.

ARC* (geometric term)

Area SUBSTANCE

ARRANGE The significant facing relation between 2=X POLARized CONFORMs or PORTIONs of 1=Y CONFORM.

ASSEMBL ASSEMBL=NT consists in adding extraneous SUBSTANCE to 1=Y unitary CONFORM.

Assembly CONCAT

ASSOCIATE ASSOCIATE=E is the condition of COMMON PORTIONs &/or ELEMENTs, if any, of 2=X CONFORMs.

cf. COMMON

INTERSECT

AXANGL( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXANGL when their LONGAXs are INTERSECT.

AXASKEW( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXASKEW when their LONGAXs are both non-parallel and non-INTERSECT.

AXBYPAS( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXBYPAS when their LONGAXs are never INTERSECT.

AXINLIN( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXINLIN when their LONGAXs are COINCIDEnt.

AXIS An imaginary STRLIN of indefinite length with respect to which the SUBSTANCE of a CONFORM is symmetrically arranged.

cf. AXANGL( )
AXASKEW( )
AXBYPAS( )
AXINLIN( )
AXMEET( )
AXOFFSET( )
LONGAX

AXMEET( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXMEET when their LONGAXs are INTERSECT.

AXOFFSET( ) 2=X CONFORMs, PORTIONs, and/or ELEMENTs are each AXOFFSET when their LONGAXs are parallel.

=B A modulating suffix for a numeral, causing it to be read as an ordinal, e.g., 3=B is read "third."

Bead MODULE

Before AFORLAP(LAPAFT)
TIMAFOR(AFTIM)

Begin SYNCSTART-(STARTSYNC)

Behavior PROCESS

BETWEEN BETWIX(BETWEEN)
STANWIX(BETWEEN)

(In the sense of a CONDITION or STATUS lasting between 2=Y other CONDITIONs or STATUSs, use either:)

SYNCSTART-(STARTSYNC)
SYNCSTOP(STOPSYNC)

(In the sense of from among, e.g., choose between:)

FROMWHENCE(WHENCEFROM).

BETWIX(BETWEEN) A BETWEEN CONFORM is BETWIX 2=Y CONFORMs when it is the center of 3=Y in a row, the 2=Y outer each being BETWIX.

Blemish ADORNBLEM

Body CONFORM
ENTITY

BORDER The boundary, limit, or periphery of a CONFORM (and thus, by definition, also a CONFORM), e.g., the end of a LINE, the edge of a surface, or the exterior surface of a volume.

cf. COMBOR( )

Boundary BORDER

Bulk SUBSTANCE

Chain CONCAT

Change PROCESS

CIRCLE* (geometric term)

Circular cylinder RTCIRCYL

Close to APPROXTIM( )

COINCIDE 2=X CONFORMs COINCIDE when the substance of all is COMMON, e.g., 2=X superimposed congruent CONFORMs.

COMBOR( ) COMMON BORDERs

COMMON The SUBSTANCE of 2=X CONFORMs, PORTIONs, or ELEMENTs occupying the same space at the same time.

cf. ASSOCIATE
COMBOR( )
COMPOR( )
INTERSECT

Component MODULE

COMPOR( ) COMMON PORTIONs

COMPRISOF(STOCKFROM) 2=X STOCKFROM MODULEs are ASSEMBLed into a COMPRISOF ENTITY when the ENTITY at least includes these MODULEs, but may include others.

CONCAT An ENTITY formed by FASTEN=NT or interengaging of a series of LINKs--MODULEs--together, LINK to LINK. The means for CONCAT=NT may be a separate MODULE, or the LINKs may have shaped PORTIONs which interengage.

Concatenate CONCAT

CONDITION A period of time (which may be instantaneous) DURING which a STATUS remains unchanging, e.g., night.

cf. STATUS

## BIBLIOGRAPHY


Andrews, Don D. and Simon M. Newman. "Activities and Objectives of the Office of Research and Development in the U. S. Patent Office." *Journal of the Patent Office Society*, 40:2 (February 1958), 79-85. (This paper includes a bibliography which lists, *inter alia*, all the Reports of the Office of Research and Development thru July of 1958.)

Andrews, Don D. and Simon M. Newman. ***Storage and Retrieval of Contents of Technical Literature, Nonchemical Information, Preliminary Report.*** Patent Office Research and Development Report [No. 1]. Washington 25, D. C., Department of Commerce, 1956. (This paper concerns the development of a linguistic approach to the search problem.)

Dunham, B. "The Formulation of Scientific Languages, Part I: The Work of Woodger and Hull." *I. B. M. Journal of Research and Development*, 1:4 (October 1957), 341-348.

Harris, Z. H. "Linguistic Transformations for Information Retrieval." ***Preprints of Papers for the International Conference on Scientific Information.*** Washington, D. C., National Academy of Sciences–National Research Council, 1958, Area V, pp. 123-136.

Locke, W. N. and A. D. Booth. ***Machine Translation of Languages.*** Boston, The Technology Press of the Massachusetts Institute of Technology, and New York, John Wiley and Sons, Inc., 1955.

Newman, Simon M. "Linguistics and Information Retrieval; Toward a Solution of the Patent Office Problem." ***Monograph Series in Linguistics and Language Studies, No. 10.*** Washington, D. C., Georgetown University Press, 1957, pp. 103-111. This paper has also been reprinted in the *Journal of the Patent Office Society*, 39:10 (October 1957), 720-729. (This paper summarizes the Office of Research and Development Reports Nos. 1 and 4.)

Newman, Simon M. ***Linguistic Problems in Mechanization of Patent Searching.*** Patent Office Research and Development Report No. 9. Washington 25, D. C., Department of Commerce, 1957. (This paper restates the language problem for the benefit of linguists.)

Newman, Simon M. ***Problems in Mechanizing the Search in Examining Patent Applications.*** Patent Office Research and Development Report [No. 3]. Washington 25, D. C., Department of Commerce, 1956. (This paper is a statement of the problems which must be resolved in mechanizing the patent search.)

Newman, Simon M. ***Storage and Retrieval of Contents of Technical Literature, Nonchemical Information, First Supplementary Report.*** Patent Office Research and Development Report No. 4. Washington 25, D. C., Department of Commerce, 1957. (This paper is a continuation of the linguistic approach described in the Office of Research and Development Report No. 1.)

***Standard Handbook of Prepositions, Conjunctions, Relative Pronouns and Adverbs.*** New York, Funk and Wagnalls, 1953.

Vickery, B. C. "The Significance of John Wilkins in the History of Bibliographical Classification." *Libri, International Library Review*, 2 (1953), 326-343.

Yngve, Victor H. "Feasibility of Machine Searching of English Texts." ***Preprints of Papers for the International Conference on Scientific Information.*** Washington, D. C., National Academy of Sciences–National Research Council, 1958, Area V, pp. 161-169.



43425--U.S.Dept.of Comm--DC--1958