A Universal Source Thesaurus as a Classification Generator

The construction of a Universal Source Thesaurus (UST) is suggested. Unlike a universal classification, UST's main purpose would not be to serve as an indexing language to be applied directly on an universal scale but to:

(a) serve as a source in the construction of special schemes;
(b) serve as a concordance between a great many classification schemes, existing ones as well as new ones.

The following problems arising in the planning for UST are dealt with briefly:

(1) The structure of UST—UST would have to be as complete and detailed as possible.
(2) The use of UST in the construction of special schemes—especially the "extraction" of special schemes.
(3) The use of UST as a concordance or "switching language" between different classification schemes as a basis for cooperative cataloguing and the possibility to have special schemes common to a group of related institutions (semi-universal classifications).
(4) Organization and management of UST.

DAGOBERT SOERGEL

Associate Professor
School of Library and Information Services
University of Maryland
College Park, Maryland

• Universal Source Thesaurus Versus Universal Classification

There is an increasing discussion, stimulated by the necessity for cooperation in information services, on the possibility of a truly universal classification. A case is made that the Universal Decimal Classification (UDC) should take this role. Others suggest a complete rebuilding of UDC. The Classification Research Group (CRG) has suggested a faceted approach without agreeing on a set of facets to be applied to a universal scheme. It is very unlikely that agreement on the principles on which a universal classification should be based will ever be reached *(1—5)*.

Three major points can be made in favor of a universal classification:

(1) It is unnecessary for each information center to construct its own classification scheme and the costs arising from such construction can therefore be saved.
(2) A universal classification enables cooperative subject indexing. Tremendous savings can thus be achieved in this very expensive area.
(3) The user needs to learn only one scheme in order to find his way through all the information services he is using.

On the other hand we have the advocates of special classification schemes. They argue as follows: Each company or each research laboratory has its own specific problems. An information service in such an environment has to be based on a classification scheme that reflects the interests and viewpoints important in this environment. Also subject indexing has to be done especially for that environment. In other words a universal classification and universal subject indexing would not help in these circumstances.

It is the main point of this article that there is no either/or in this question; both sides are right to some degree. But how can both viewpoints be reconciled? Let us look how a similar problem has been solved in an entirely different field, the field of computer programming.

In computer programming we have a number of tasks which are recurrent in many situations, for example tasks in accounting. One could therefore be tempted to think that a program for such a task could be written once and then be applied in any situation where it occurs, thus avoiding unnecessary duplication in programming effort. However, a closer look quickly shows that this would not

be feasible: whereas the task is basically the same in different situations, there are variations in detail and there are variations in the number of data to be processed. Then the makes and types of computers used vary and a program written for one installation can not easily be used for another. These factors seem to compel programmers to write a new program from scratch in every instance. However, this is not what happens either. Programmers have found intermediate solutions between the two extremes. There are three basic types of such solutions (the dividing lines being not sharp, however).

(1) The development of higher level programming languages. In a higher level programming language recurrent small programming tasks have been preprogrammed so that the programmer instead of writing the same sequence of assembler instructions again and again can just use one statement in the higher order programming language to replace that sequence. Also, with certain limitations, a program written in a higher order language can be run on different makes and types of computers.

(2) Development of subroutines or program modules. Program modules are written for small, well defined tasks that occur in the framework of many different larger programs. They are usually built in such a way that they can be adapted to specific circumstances by the use of parameters. A programmer who encounters a specific task in the framework of a larger program has not to program this task himself but he can use an available module, he has only to specify the appropriate parameters. Many programming tasks are thus reduced to putting together available building blocks.

(3) In addition, there are program packages for major tasks, where again a range of variation is possible by the use of parameters. If this available package is suitable for the computer at hand and if it otherwise meets all the requirements then no new program has to be written at all.

In conclusion we may say that in the field of programming the emphasis is not on the development of universal programs but on the development of programming aids which make it much easier to write specific programs for specific situations. The same principle should be applied in the development of classification schemes. Instead of trying the unfeasible, namely developing a universal classification scheme, we should concentrate our effort on the feasible, developing aids for the construction of special classification schemes *{6, 7)*. It is suggested in this paper to develop a Universal Source Thesaurus (UST) for that purpose. As described below such a universal thesaurus would contain a vast amount of (multilingual) terminological and classificatory information in machine readable form, ready to use for anybody who wants to develop a special classification scheme.

Up to now we have been concerned with the development of classification schemes (point (1)). The second and perhaps more important argument in favor of a universal classification is that it enables cooperation in subject indexing. Let us see what UST could do in this area. It has been observed by many writers that the development of "concordances" between different schemes is a more realistic approach to cooperation in subject indexing than the development of a universal classification. (This does not exclude the possibility that a specific community of information centers and libraries may use the same classification scheme in order to facilitate cooperative subject indexing even more.) In particular this is true with view to the many schemes in existence and in heavy use. A properly constructed UST would serve as such a concordance not just between two schemes but from almost every classification scheme or other indexing language into any other. By its very use in the construction of new special classifications it would make sure that these classifications are included in the concordance.

The following points have to be considered in persuing this project of a universal source thesaurus further:

(1) The structure of a universal thesaurus.
(2) The use of a universal thesaurus in the construction of special schemes.
(3) The use of a universal thesaurus as a concordance or "switching language"[7] between different classification schemes as a basis for cooperative cataloguing.
(4) Organization and management of a universal source thesaurus.

1. The Structure of A Universal Source Thesaurus

1.1 The universal thesaurus would not adhere to any single principle of classification. On the contrary it would be organized in such a way that every hierarchical relationship contained in any other classification scheme (or suggested by any serious user) can be included, and information concerning the overall arrangement according to any principle of classification can be stored. In this way a follower of faceted classification may have printed for him a list of the concepts relevant to a certain field or all concepts contained in the thesaurus in form of a faceted classification. Somebody who prefers an arrangement according to the lines of scientific disciplines may have a printout this way. Or one may printout an alphabetical list of subject headings. Only by this neutrality with regards to the principles of classification is it possible to achieve universality. On the other hand a universal thesaurus would of course make use of the insights gained by modern classification theory. The principle of polyhierarchy has already been referred to in the previous paragraph. Compound concepts would be broken downinto semantic factors in order to facilitate the study of
their interrelationships. For further elaboration see ***(S,9).***

1.2 A universal thesaurus would contain concepts down to the very deepest level of specificity useful for any specialized library or information center in any field.

Otherwise it would not be possible to extract specialized schemes, or to use the universal thesaurus as a concordance. For each term very detailed information would be given. E.g. synonyms would include translations into a variety of languages. In a narrower term—broader term relationship there would be noted whether it is a generic relationship or a part-whole relationship. Special care would be taken in the addition of scope notes and references to works where the concept is defined and explained in more detail. This information will be of tremendous value in constructing a specialized scheme, as anybody who was involved in such work may acknowledge.

1.3 In the construction of a universal thesaurus a great number of existing classification schemes and thesauri would be used as sources *(10)*. It would be a "cumulative thesaurus" containing all the information contained in any of the sources. Procedures for the production of such a cumulative thesaurus with the aid of computers are described in *(9)*. These procedures in particular take into account the function of a cumulative thesaurus as a concordance between the classification schemes and thesauri used as sources.

1.4 In the production of the universal thesaurus a "guidance classification scheme" could be developed. This scheme would reflect the best knowledge of the full-time specialists working on the universal thesaurus and all the subject experts asked for advice and review. In fact two such guidance classification schemes would be developed, one arranged according to the principles of faceted classification, the other arranged according to more traditional lines. These guidance classification schemes would, however, in no way be compulsory, they would represent merely suggestions to people constructing specialized schemes.

2. Use of A Universal Source Thesaurus in the Construction of Special Classification Schemes

We have emphasized the role of a universal thesaurus as a source of information in constructing specialized schemes (as opposed to its use as a universal classification scheme). That's why we call it a universal *source* thesaurus.

There are two methods to use the universal source thesaurus in the construction of a special scheme. The first, called the "extraction and-modification method" makes use of the full information contained in the universal thesaurus, including the guidance classification. The second uses the universal thesaurus strictly as a collection of terms and other information from various sources, thus saving the clerical work involved in the first steps of the construction of a classification scheme of a thesaurus.

2.1 The Extraction-and-Modification Method

In this method, the guidance classification scheme (in either the faceted or the more traditional arrangement) is first printed out in whole or in selected parts. Since, as we have noted above, concepts up to a high level of specificity are included in the universal thesaurus it would seldom be practical to print the whole guidance classification scheme. One may print out for example "the part electrical engineering and concepts from other fields important for electrical engineering, in faceted arrangement". Or one may do the same for the field of psychology. One may also give a command "omit very specific concepts" (very specific concepts are marked as such in the universal thesaurus, and there is no absolutely fixed relationship between hierarchical level and being very specific.)

Based on this result an "extraction specification" is written for the computer, using a simple format described m section 2.1.1. Note that any modification desired can be included in this specification: the order of facets or of fields or the order of subfields within a field or the order of coordinate concepts within an array may be changed. Concepts may be deleted or new concepts may be added. All these changes are very easy due to the format used in writing the specification. The major question is of course what modifications should be made in the first place. This has to be determined by a careful study of user needs, by collection of actual and expected search requests and by discussion of the scheme with the users.

The extraction specification is then processed by the computer, resulting in the following products:

(a) The schedules as specified in the language specified. A tentative notation reflecting the specified arrangement has been assigned by the computer program.
(b) The main part of the specialized thesaurus, containing for each term all information available in the universal thesaurus, such as synonymous and equivalent terms, additional broader and narrower terms, scope note.
(c) An alphabetical index.

The main part should now be studied in detail in order to determine whether the information contained in it is agreeable and sufficient. Again, all desired modifications can be entered in a specialized format. The schedules should also be checked because it is now easier to see whether they really reflect the needs of the specific environment.

The modifications are entered into the computer which there upon produces final versions of (a)—(c).

2.1.1. Format for "Extraction Specifications"

The format is explained using an example which is simple yet elaborate enough to show the potentialities of the method. The guidance classification" is given in fig. 1. Two extraction specifications and the resulting schedules are shown in fig. 2 and fig. 3. The corresponding main part and alphabetical index is shown in neither case. It is to be expected that in any real application further

problems would arise and lead to more elaborate rules for the extraction specifications. (The format in a slightly different version, has first been described in *(11)*.

2.2 The method described in the previous section relies heavily on the guidance classification scheme. If one does not wish to do so one may still make use of the universal thesaurus as a collection of terms and other information from a large number of sources. The first steps in a construction of a classification scheme or a thesaurus are, as is well known, the collection of terms and information about their relationships and the elimination of duplicates. One important source for terms is other classification schemes and thesauri. Other important sources include titles, abstracts, indexes of textbooks, studies of user needs and search requests. The collection of terms

WU	DOCUMENT REPRODUCTION
WU2	*Photocopying*
WU3	Basic photocopying techniques
WU4	Direct photocopying
WU5	Reflex photocopying
WUA	Photographic photocopying
WUB	Silver halide processes
WUJ	Diazo process
WUL	Thermography
WUN	Adherography
WUQ	Dual spectrum process
WUR	Infrared reflex process
WUS	Infrared transfer process
WUT	Infrared sublimation process
WUV	Electrostatic copying
WUX	Xerography
WUY	Electrofax
WV	*Microphotography*
WV2	Microforms
WV3	Flat microforms
WV4	Microtransparencies
WV9	Microopaques
WVD	Roll microforms
WVE	Microfilm
WVG	Microform equipment
WVH	Microfilm cameras
WVHB	Planetary microfilm cameras
WVHE	Rotary microfilm cameras
WVHM	Step and repeat microfilm cameras
WVI	Microform processors
WVJ	Microform mounters
WV	Microform printers
WVM	Microform readers
WVQ	Micropublishing
WVS	*Duplicating*
WVT	Carbon copy duplicating
WVU	Stencil duplicating
WVV	Typing stencils
WVV1	Writing on stencils
WVV2	Electronic stencil cutting
WVW	Spirit duplication

FIG. 1. The guidance classification scheme *

* Taken from Thesaurofacet, with modifications.

Extraction specification

	Comment:
WU#111	Take WU itself and all the descriptors being one or two levels below (each 1 after # stands for one level, the first level being WU itself)
*NOTATION TYPE1 USE KOO-K 99	Assign notation of type 1, using range KOO-K 99

Resulting schedule with tentative notation

KOO DOCUMENT REPRODUCTION

K30 *Photocopying* K32 Basic photocopying techniques K34 Photographic photocoping K36 Thermography K38 Electrostatic copying

K50 *Microphotography* K53 Microforms K55 Microform equipment K57 Micropublishing

K70 *Duplicating* K73 Carbon copy duplicating K75 Stencil duplicating K77 Spirit duplication

FIG. 2. Simple extraction specification and resulting schedule

and the elimination of duplicates involve a heavy clerical effort. Part of this effort can be avoided by using the universal thesaurus as a cumulative thesaurus of those sources that one intends to use and that have been processed in the construction of the universal thesaurus. One may simply request an alphabetical list of the terms contained in the specified sources, accompanied by the further information on these terms taken from these sources. These terms may even be printed on thesaurus forms for the convenience of further processing.

3. The Use of a Universal Thesaurus As a Concordance or "Switching Language" Between Different Classification Schemes in Cooperative Subject Indexing

Once the concepts for which a document is relevant have been determined and expressed by a term or a notational symbol from one classification scheme, it is a purely mechanical matter to look up the term or notational symbol to be used with other classification schemes, provided the concept is available in the other classification scheme. For cooperation in subject indexing the major requirement is therefore conceptual compatibility between the classification schemes used in the cooperating libraries/information centers. The arrangement of concepts in these classification schemes and whether no-

Extraction specification

Instruction No.		Comments Take
1	WU#1	WU itself
2	WVS#11 (WU,01)	Starting one level below WU (therefore '(WU,01)') take WVS and all descriptors one level below
3	WU2#111	Starting on the same level as WVS, take WU2 with all descriptors one or two levels below
4	(/WUBa, 01/	( ) signifies a modification. Between / / the place is given. Here something has to be inserted after WUB (therefore 'WUBa'), starting one level below WUB (therefore',01')
	Photostat process Photostabilization process Direct positive process Diffusion transfer process Gelatin transfer process)	This is the block to be inserted
	(/WUN-WTJT/==)	Take the block starting with WUN and ending with WUT and eliminate it (= = stands for a block containing nothing, replacing the original block)
6	WV#2	Starting on the same level as WU2 take WV and all narrower descriptors ('#2')
7	(/WV3-WVE, 1/	Take the block WV2-WVE and replace it by the following block, starting on the level of WV3 (therefore, ',1')
	Micro transparencies Microfilm (roll) Flat microtransparencies Microfiches Aperture cards Film strips Microopaques)	Replacing block. Note that indentations signifying hierarchy are allowed in writing blocks
	*NOTATION TYPE 1 USE KOO-K99	Assign notation on type 1, using range KOO-K99

Resulting schedule

KOO DOCUMENT REPRODUCTION

K20 *Duplicating*
K23 Carbon copy duplicating K25 Stencil duplicating K27 Spirit duplication — Note that this division has been moved from the bottom to the top (instruction no. 2)

K30 *Photocopying*
K35 Basic photocopying techniques
K37 Direct photocopying K38 Reflex photocopying

K40 Photographic photocopying
K42 Silver halid processes
K42.2 Photostat process K42.4 Photostabilization process K42.5 Direct positive process K42.7 Diffusion transfer process K42.8 Gelatin transfer process — Insert (instruction no. 4)
K44 Diazo process

(Fig. 3 continued on next page)

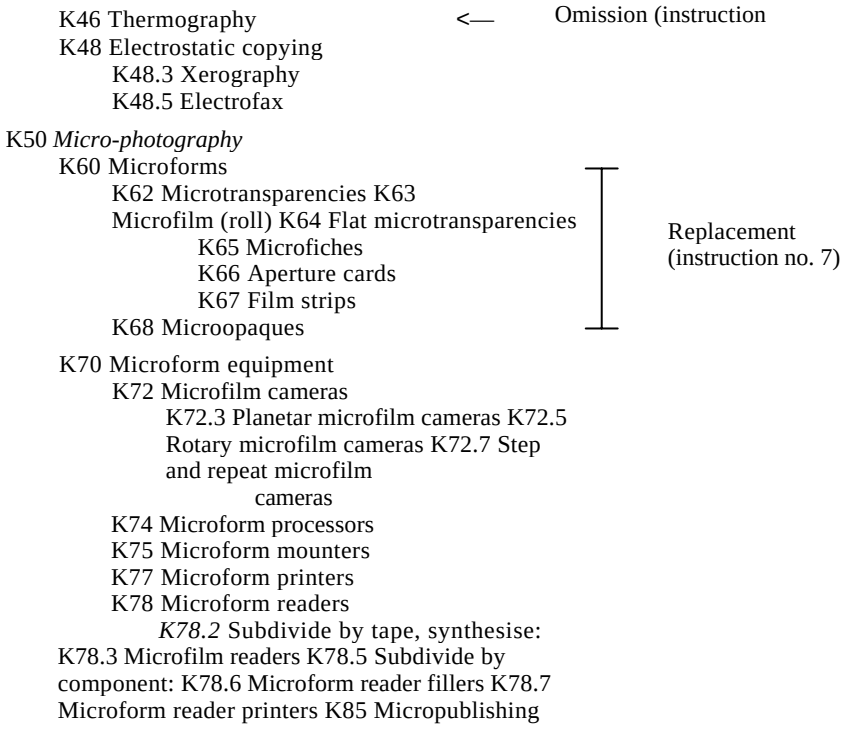

FIG. 3. More elaborate extraction specification and resulting schedule

tations or natural language terms are used is of very minor importance. In other words, cooperative subject indexing does not necessitate a universal classification; conceptual compatibility between different schemes and the availability of a concordance between them are sufficient.

This statement has to be qualified, however, by two major caveats:

(1) Classification, by displaying the viewpoints to be considered in indexing, plays itself a major role in the determination of the concepts for which **a** document is relevant. As is well known this is one of the major reasons why special classification schemes are needed and one of the major reasons militating against cooperation in subject indexing. This is the essence of the "filtering technique" of indexing *{12}*.

(2) There are limiting technical factors. If two co-operating libraries use computer systems then it is very easy to substitute the notational symbol or term used in library A for indexing by the notational symbol or term used in the catalog of library B. If we have a medium size library using conventional methods and receiving Library of Congress catalog cards and/or books that are already labeled by the call number according to Library of Congress Classification, then the situation is very different and cooperation in subject indexing requires indeed a common classification scheme.

This leads us to another point. The information needs of the clientele of public libraries or academic libraries or school libraries are fairly similar within each type, at least within the same type of community. Therefore there is no good reason why each type of library should not apply a common classification scheme and it would even be of advantage if the schemes for different types of libraries would have a common basis. The same is true for the libraries and information systems of, say, the weather bureau around the world or any other group of institutions with similar purposes. Common "semi-universal" classification schemes could be constructed in these cases so as to accommodate the viewpoints important for any of the institutions in the group, thus taking care of caveat *(1)*. The creation of common semi-universal classification schemes for thematic groups of institutions could thus completely facilitate the cooperation of subject indexing within those groups—and that is in any case where the bulk of cooperation would take place.

Of course in each case a major discussion would arise as to what principles should be at the basis of the common, semi-universal classification scheme. It seems to this writer, however, that this problem could be solved much easier in a limited context than in the broad context of a universal classification.

In summary, then, it is proposed to create a series of classification schemes, each common to a group of institu-

tions facing a similar situation. The scope of these schemes may be either universal (as in the case of academic libraries) or specific (as in the case of the weather bureau). They would all be incorporated (along with all the additional special schemes used) in the universal source thesaurus.

Because the universal source thesaurus would be used in the construction of new special schemes it would contribute to an increased compatibility of and similarity between new schemes. And, of course, the quality of new schemes would be increased.

The universal source thesaurus would also be used in the modification and updating of schemes already in existence and thereby contribute to an increasing compatibility between these schemes, too.

4. Organization and Management of A Universal Source Thesaurus

In the previous sections the enormous benefits to be derived from a UST have been described. There should be no illusions that a price has to be paid in order to achieve these benefits. The construction and maintenance of a UST would require considerable organizational background. First of all a central organization would be needed, staffed by classification theorists, subject experts in the different disciplines, systems analysts, computer programmers, keypunch operations, and other clerical staff (one could easily imagine a staff of 250 people) and equipped with sufficient computer facilities. In addition there should be regional offices in all parts of the world which would make the services available to libraries/ information centers. The same organization would take care of the administration of the various common classification schemes mentioned above.

When all existing classification schemes and thesauri have been included one has to keep track of the new schemes being developed. There are now the centers in Cleveland, Ohio and Warsava trying to do this. However, the incentives are not very great. With UST and the organization backing it things would be very different. Anybody building a new scheme would want to use UST and its services because he would save much effort in doing so. Thus UST would be informed of nearly all projects of constructing a new special classification and, what is more, since UST programs and computers would be used in many of these projects the new classification scheme or thesaurus would automatically be included in UST.

If adequate financial means and organizational backup are not available one may just as well forget about UST. The current apparatus of FID, for example, would be utterly inadequate for such a task. This does not exclude projects that would create something along the lines of UST for a specific field and/or a specific group of institutions and/or with a limited number of existing classification schemes as sources. Austin's work at BNB on PRECIS, for example, might well result in a cumulative thesaurus linking Library of Congress Classification, Dewey Decimal Classification, and a faceted classification developed in the framework of PRECIS *(13)*.

References

1. Classification Research Group (Ed.), *Classification and Information Control,* The Library Association, London, 1969, 230 p.
2. Classification Research Group, Classification Research Group. Bulletin No. 9, Journal Documentation v. 24, no. 4 (1968.12) p. 273-298 (Dec. 1968).
3. Foskett, D. J., *Classification for a General Indexing Language,* The Library Association, London, 1970.
4. AUSTIN, D., *Prospects for a New General Classification,* Journal of Librarianship, 1 (No. 3): 149-169 (1969).
5. DAHLBERG, I., *Principles for the Construction of a Universal Classification System.*—A Proposal, paper presented at the 1. Ottawa Conference on the Conceptual basis of the Classification of Knowledge, Ottawa. October 1970.
6. KYLE, B., *Classification: Adopt, Adapt, or Create? A Discussion Point,* Aslib Proceedings, 12 (No. 9): 317-320 (Sept. 1960).
7. KLINGBIEL, EL, *The DDC Thesaurus: Past, Present and Future,* Aslib Proceedings, 16 (No. 8): 252-257 (Aug. 1964).
8. SOERGEL, D., *A General Model for Indexing Languages: The Basis for Compatibility and Integration,* in Wellisch, H., et. al. (Ed.), Subject Retrieval in the Seventies, Westport, Conn., Greenwood, 1972, pp. 36-61.
9. SOERGEL, D., *Construction and Maintenance of Indexing Languages and Thesauri,* in preparation. This book will contain an extensive bibliography.
10. MOOERS, C. N., *The Next Twenty Years in Information Retrieval.—Goals and Predictions,* Proceedings of the Western Joint Computer Conference, San Francisco, Calif. 15: 81-86 (March 1959).
11. SOERGEL, D.. and K. FURMANIAK, *The Description of Statistical Tables—A Problem in Data Documentation,* Proceedings of the American Society for Information Science 7: 331-336 (Oct. 1970).
12. MOOERS, C. N., and C. W. BRENNER, *A Case History of a Zatocoding IR System, in* Casey, R. S., et al. (Ed.), Punched Cards, 2 ed., Reinhold, New York, 1958. pp. 340-56.
13. AUSTIN, D., *The PRECIS System for Computer-generated Indexes and its Use in the British National Bibliography,* in Wellisch, H., et al. (Ed.), Subject Retrieval in the Seventies. Westport, Conn., Greenwood 1972. p. 99-115.