Part III

Procedures for the Construction and Maintenance of Indexing Languages and Thesauri

Chapter F

Flow of Work in the Construction of Indexing Languages and Thesauri

F0 OVERVIEW AND GENERAL PROBLEMS

F0.1 The Major Steps (See Overview Flowchart, Figure 52)

The logical sequence of major steps is as follows: The flow of work naturally starts with the collection of material (F1). (Many ISAR systems start without a thesaurus, and terms chosen freely by the indexer are used in indexing. This may be considered the collection phase. Unless the ISAR system is very small, one will soon detect that this approach is insufficient and then proceed with the construction of a thesaurus based on the terms collected in indexing using free terms.) Thereupon follow the steps F2 to F5. Step F2, "Sort into alphabetical order and merge information on identical terms", consists of a series of purely clerical procedures which are geared to reduce the redundancy found in the material as originally collected. This frees the editor or lexicographer to concentrate on conceptual work in Steps F3 to F5. Step F3, "Work out the structure of the thesaurus", is mainly a "process of distillation": from among all the terms collected the preferred terms are selected. The preferred terms are in turn examined to determine whether or not they should be selected as descriptors. The concepts designated by the preferred terms are decomposed into semantic factors as far as possible. In this step the decomposition into semantic factors serves mainly the purpose of detecting elemental concepts to be included in the indexing language. The result of this distillation process is a listing of the more important preferred terms, which contains, in condensed and easily comprehensible form, most of the information that has been collected in step F1.

The result of Step F3 is a first draft of the hierarchical arrangement of the descriptors and other important preferred terms. In Step F4, "Work out

first draft of the classified index", the hierarchical structure is elaborated and improved upon. Since only the more important preferred terms are retained for this step, it is possible to get a picture of the whole and to streamline the overall structure. After this step the design of the indexing language and of its structure is essentially finished. On this basis it is now possible to elaborate in Step F5, "Complete first draft of the thesaurus as a whole", the complete structure of the thesaurus, making use of the work that has already been done in Step F3. One might say that one returns from the condensed classified index to the complete mass of material.

The result of Step F5 is a first draft of the whole thesaurus. Before the thesaurus is actually used, however, it should be put to a practical test in indexing and retrieval experiments (Step F6). From these a number of modifications may arise. Only after this step is the thesaurus ready for distribution and actual use. Of course there will be further additions and modifications that come up during the use of the thesaurus. This problem is dealt with in Chapter J, "Updating and maintenance of indexing languages and thesauri".

Figure 53 gives a detailed flowchart for later reference.

F0.2 Cooperative Thesaurus Development

In many cases it is useful for a group of institutions to share efforts in thesaurus development. This cooperation may be strictly for the purpose of saving effort in such thesaurus development or with a view to sharing the results of subject indexing. These topics are dealt with in Chapter K, especially Sections K1.2, "Cooperation in the development of the terminological and classificatory structure" and K2.3.1, "Production of conversion tables. Ideal situation: the indexing languages of the cooperating institutions have yet to be built".

F0.3 Collaboration of Experts from Different Subject Areas

F0.3.0 Necessity of full-time staff and collaboration of subject experts

Thesaurus construction is *not* a task appropriate for a committee of subject experts alone; it requires an adequate staff. On the other hand, a good thesaurus can only be developed in close collaboration with experts in the appropriate subject areas and specialized fields. It is of advantage if some of these subject experts have experience in indexing documents and in searching ISAR systems and if others come to the task without having their thinking constrained by the practical limitations inherent in any ISAR system.

There are various organizational frameworks that can be used in the consultation of experts. If the thesaurus is developed for a specific institu-

tion, it may be sufficient to consult experts from only that institution. If the thesaurus is developed for a broader network or even as a national or international thesaurus, the range of experts has to be broadened accordingly. In either case one may consult individual experts (F0.3.1 and F0.3.2) and/or set up a structure of committees or panels (F0.3.3): a central or full committee for making the ultimate decisions and any number of panels to deal with specific subject fields or subfields. (There may even be a further hierarchy among these panels. Such a structure has been used in building TEST, and it is also used for updating UDC. In the development of the EJC Thesaurus (1st edition) ten committees for fields such as *Electrical and Electronic* or *Aerospace* were set up. An elaborate hierarchy of panels has been set up for the updating of UDC.)

Subject expertise is especially important for the basic decisions in developing the classificatory structure. The consultation of subject experts should therefore be especially intensive in the determination of the major areas in which the indexing language is to be divided (Steps F3.1, "Sort terms into broad subject fields"), in the elaboration and streamlining of the classificatory structure (Step F4, "Work out first draft of the classified index"), and in the review of the whole thesaurus (Step F5.5).

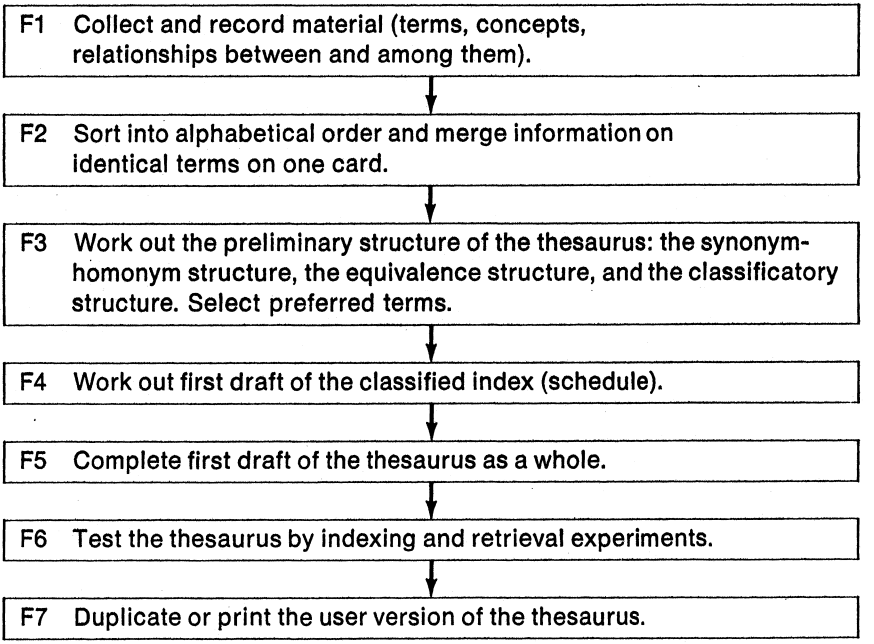


Figure 52. Flow of work in thesaurus construction: overview flowchart (F0.1).

Figure 53. Flow of work in thesaurus construction: detailed flowchart (F0.1).

Step #	To be performed by	See also Section	Work to be done
F1.1	Professional staff	F8.5	Select sources for the collection of terms, etc.
F1.2.1,0	Semiprofessional staff		Assign an abbreviated code to each source.
F1.2.1,1 F1.2.1,2	Professional staff Semiprofessional staff		Selection of terms (for part of the prearranged sources, for all open-ended sources). Assignment of an auxiliary notation (for some prearranged sources).
F1.2.2 or	Clerical staff Clerical staff, keying device	D2; F0.5; F1.2.3; G1.2.2	Transfer selected terms with all information on thesaurus forms. Punch selected terms with all information.
F2.1	Clerical staff or computer	G2	Sort into alphabetical order.
F2.2	Semiprofessional staff or computer	F0.7.2; F2.4; G2; G2.2.3; G2.4	First round of merging: merge information for identical terms. Possibly "pulling" information from additional sources.
F2.3	Semiprofessional staff or computer	F0.7.2; F2.4 G2.3; G2.4	Second round of merging: merge information or terms in the same concept class. (Computer: printout on thesaurus forms)
F3.1	Professional, semiprofessional staff, or computer	C1.4.3; F0.3.2; F0.3.3; G3 excl. subsect.	Define broad subject fields and sort terms into these broad fields. (Computer: printout on thesaurus forms)

(Continued)

Step #	To be performed by	See also Section	Work to be done
F3.2	(Subject experts), professional, semiprofessional staff, ((or computer))	F0.3.2; G3 excl. subsect.	Define subfields within one broad subject field and sort terms into these subfields. (Computer: printout on thesaurus forms)
F3.3	(Subject experts), professional, semiprofessional staff, ((or computer))	B; C; F0.3.2; F0.4; F0.7; F0.8; F8.4; F8.5; G3.3; G3.4	Work out the detailed structure. Select preferred terms. Merge information for terms in the same concept class. (Computer: printout on thesaurus forms; then manual editing.)

All subfields of broad subject field finished?

No

Same for next subfield

Yes

All broad subject fields finished?

No

Same for next broad subject field

Yes

F8.1

F4.1	Clerical staff or computer	D3, esp. D3.1.1 G0.2(4) and (5)	Type preliminary version of classified index. Amend working file.
F4.2	Professional with semiprofessional staff or computer	B5; C	Improve classificatory structure.
F4.3	Clerical staff (computer)	F4.0; G4	Type improved version of classified index. Distribute among subject experts. Amend working file.
F4.4	Subject experts, professional, semiprofessional staff	B5; C; F0.3.3; F0.4	Discuss classified index with subject experts. Select descriptors and checklist descriptors.

Many modifications?

Yes

No

(Continued)

Step #	To be performed by	See also Section	Work to be done
F4.5	Semiprofessional (professional), or computer	D4; F8.2; G0.2(5)	Assign notational symbols.
F4.6	professional		Make a systematic search for additional cross-references.
F5.1	professional with semiprofessional or computer	B5; B6; C; D2 E0; E1; E2; F0.7; F0.8; F3.3.2; F4.0; G5.1	Revise all entries in the working file as follows: (a) formulate standardized abbreviation; (b) standardize form of Main Term; (c) standardize elements in BT, NT, RT to consist of notation and preferred term; (d) improve classificatory structure; (e) USE, UF, etc.
F5.2	Clerical	D2; F5.10.2; F0.8.3; F9; G0.2(6)	Produce the main part of the thesaurus in list-form.
F5.3	Semiprofessional, clerical or computer	C7; G5.3	Check inverse cross-references and insert where necessary.
F5.4	Clerical		Insert modifications in the manuscript prepared in F5.2, duplicate and distribute among subject experts.
F5.5	Subject experts, professional/ semiprofessional	F0.3.3	Review the whole thesaurus. Consult with subject experts.
F5.6	Semiprofessional with clerical or computer		Insert modifications; repeat F5.3.
F5.7	Semiprofessional with clerical or computer	E3 F9	Produce alphabetical index. Test model: produce alphabetical main part and alphabetical index.
F5.8	Professional/ semiprofessional		Check homonyms and cross-references using alphabetical index.
F5.9	Clerical	F5.10.2; F9; G7	Reproduce test version of thesaurus.
F6	Subject experts		Test the thesaurus by indexing and retrieval experiments. Insert modifications.
F7	Professional semiprofessional clerical or computer	F5.10.2; F9; G7	Duplicate or print the user version of the thesaurus.

End

It is important that the consultation of subject experts be carefully pre-
red by the staff so that optimal use is made of the subject experts and so
it subject experts are not annoyed by trivial, unclear, or imprecise ques-
ns. Do not ask subject experts to answer questions that could be just as
sily answered by looking in a dictionary or that otherwise do not really re-
ire subject expertise. Formulate questions as precisely as possible. The
oblems to be emphasized in the collection of information from subject
perts are indicated in the description of each step of the procedure for
esaurus development.

As mentioned above, the contributions of subject experts can take three
rms, which will now be discussed in detail.

).3.1 Supply of material

(1) *Search requests and indexed documents.* Subject experts should be asked
supply search requests and/or to index documents using free terms. Even bet-
r is the soliciting of points of interest in a discussion session, using documents
; initial stimuli, to be described in Section F1.1.2(n). If this is not possible, in-
exing guidelines should be given to the subject experts. In these guidelines the
iecklist technique of indexing should be explained briefly, and the subject ex-
erts should be asked to think of all aspects for which the document might be
elevant and to use appropriate terms in indexing. The guidelines should not con-
ain rules for the form of terms or other technical details that would only dis-
ract from the major task. These details can be taken care of by the editorial staff.

(2) *Comments on draft of thesaurus.* Subject experts can be asked to give
vritten additions to and modifications of a draft edition of the thesaurus. The
easiest way for the experts to give these comments is to write them into a copy
of the draft thesaurus and return this copy. Usually these written comments yield
ewer suggestions from any one subject expert than the discussions described in
he following sections. However, this procedure has the advantage of making it
possible to secure comments from many experts.

F0.3.2 Answering questions on single problems that come up during the work on the thesaurus

Getting information from experts by asking specific questions is especially useful for the sorting of terms into broad subject fields (Step F3.1), for the individual decisions on the terminological and the classificatory structure to be made in Step F3.3, and while working out the final thesaurus structure in Step F5.1.

If possible one should not bother a subject expert with every individual question as it comes up but wait until a group of questions has been collected. In order to get different opinions it is often advisable to ask the same question of different experts in the appropriate subject fields. Questions should be asked of an expert personally or by phone, if possible, since then

a clarifying dialogue can take place. If this is not possible, a query-and-reply slip can be sent to the expert. For procedural details see Section F0.8.1,2.

F0.3.3 Discussion sessions for review and/or decisions on difficult problems

Discussion sessions are appropriate for the Steps F4.4, "Discussion of the classified index" and F5.5, "Review of the whole thesaurus" (and possibly for the definition of broad subject fields in Step F3.1). There are two types of discussions: "routine" discussions of the whole thesaurus and special sessions on difficult problems. The material to be discussed should be sent to the participants well in advance. (On keeping minutes see Section F0.8.1,3, "keeping track of why decisions have been made".)

(a) *Routine discussions of the whole thesaurus.* For most problems, a meeting of three to five people (2-3 subject experts, 1-2 information specialists) to discuss a certain subject field will give the best results. There are two possibilities for conducting such small group discussions:

(a1) Discuss selected problems only. These may be problems in which further clarification seems necessary to the staff responsible for the development of the thesaurus, or they might be cases in which at least one subject expert has indicated dissent from a draft version of the thesaurus.

(a2) Discuss descriptor by descriptor. Experience has shown that, in the discussion of a single descriptor or of the hierarchical arrangement in a specific area, useful revisions arise which none of the participants alone would have thought of.

If it is possible in terms of time and personnel, the discussion of a subject field should be repeated with a second group of subject experts (the information specialists being the same as in the first group). In most cases the results of both groups will agree. Many remaining differences may be resolved by a second discussion with the first group.

In developing a cumulative thesaurus that is to represent the viewpoints of different institutions the whole thesaurus should be discussed with representatives of each institution, possibly in two rounds, to make sure that the interests of each institution are represented adequately. The differences still remaining have to be resolved in a meeting of representatives from all institutions, as described in (b).

(b) *Special sessions on difficult problems.* The differences still remaining are probably major problems in their specific disciplines or reflect differences in the approach taken by different institutions. These differences have to be discussed and decided upon in a larger meeting. Such discussions in a larger group require especially careful preparation. The alternatives have to be worked out and presented very clearly, and the necessary documents should be sent to the participants well in advance. In no case should the whole indexing language or thesaurus be discussed in a larger group. The many detailed questions that are involved in such a discussion cannot be dealt with thoroughly and leisurely enough, and many

ill-formed decisions will be made in a hurry. In particular, it doesn't make any sense to go through an alphabetical listing, term for term, in a larger meeting.

F0.3.4 Inter-disciplinary approach

Subject experts should also be asked to look at parts of the indexing language not falling into their specialty, at least at those parts in the neighborhood of their specialty. In particular, the indexing language as presented in the classified index should be checked in its entirety by subject experts from many different fields. It might well be that a scientist can make useful suggestions in fields that are not his own specialty because he, from the viewpoint of his specialty, may bring in aspects that have been overlooked by the experts in the field in question. This is of particular importance for the transfer of information among different disciplines. There is another advantage to this procedure: it enables one to make sure that all areas of the indexing language can be understood by all subject experts to be served by the information system, regardless of their specialty. If parts of the indexing language are comprehensible only to specialists in those particular areas, discussions with subject experts should reveal how this shortcoming can be corrected. This aspect is especially important in the discussion of the classified index (Step F4.4).

F0.3.5 Briefing of subject experts on thesaurus functions

At least the subject experts consulted regularly and/or involved in discussion sessions should have some understanding of what a thesaurus is and how it is structured. The appropriate information may be given in written form (e.g., the introduction to the thesaurus) or through briefings.

F0.3.6 Source codes for subject experts and panels

Subject experts and panels consulted should be treated as sources and assigned a code symbol accordingly, as described in Sections F0.7.5 and F1.2.1,0.

F0.4 Criteria for the Selection of Terms and Descriptors

Selection processes take place in all the steps to be described later. Selection decisions are concerned with terminological problems on the one hand and conceptual problems on the other.

In Step F1, "Collection and recording of material", only those terms should be eliminated that obviously do not fall within the scope of the thesaurus. Final selection can be made only in the framework of the classificatory structure to be developed.

There are three different kinds of selection decisions, and one should

be careful to keep them separate. First, one has to decide what terms should be included at all, if only as nonpreferred synonyms (F0.4.1). Second, one has to select a preferred term from each class of synonyms and quasi-synonyms (F0.4.2). Third, one has to select the preferred terms to be included in the thesaurus and, more important, the descriptors from among the preferred terms (F0.4.3). Most guidelines on thesaurus construction confuse selecting a preferred term from a class of synonyms and selecting descriptors; they then offer a mix between two sets of criteria that should be kept separate. (It is interesting to compare the problem of vocabulary selection for a thesaurus, to be discussed in this section, with the problem of vocabulary selection in foreign language instruction. Teachers and textbook authors have relied heavily on frequency listings to determine the vocabulary that should be taught. However, it has been argued recently that semantic considerations should take precedence over the mechanical application of frequency as a selection criterion. This is exactly the point brought out in the following discussion.)

F0.4.1 Criteria for the selection of terms (whether nonpreferred lead-in terms, preferred lead-in terms, or descriptors) to be included in the thesaurus

Include every term designating any concept that is included in the thesaurus, even if the term is outdated or seldom used or in an area marginal to the field of the thesaurus. The reason for establishing this rule is that it will make the alphabetical index more useful. The time needed to look up a term in the alphabetical index does *not* increase substantially with the number of entries contained in the index (provided the index is arranged properly). On the other hand, the average time needed to find a term in the alphabetical index decreases as the probability of the term being contained in the index increases. (TEST stipulates: "Slang, jargon, coined terms, and deprecated terminology should be excluded." However, this is not a useful restriction; on the contrary, if such terms are in current use among the user group to be served, they should be included.) On the selection of spelling variants to be included, see Section C6.2.2.

F0.4.2 Criteria for the selection of a preferred term from a class of synonyms and quasi-synonyms (arranged according to decreasing priority)

The preferred term should:

(a) be the best to reflect the meaning of the concept;

(b) be recognized in the "user community". (In science and technology this usually means recognition in the national and, if possible, international scientific

community. The term should also reflect exact scientific usage and the newest terminology of the field in question. It is to be hoped that all these criteria converge.) ("The acceptability of terms can be determined by consulting dictionaries, encyclopedias, or other indexing vocabularies, and the opinions of subject specialists.");

(c) be unambiguous (not a homonym);

(d) be simple and short in spelling.

Statistics gathered in the following way may assist in the selection of the appropriate term:

(e) Identify all sources, such as other thesauri, that mention the concept in question. Some of these sources may contain several terms to designate the concept but usually one of them will be used as the preferred term. Find the term that is used by most sources as the preferred term. The strength of the terminological consensus can be measured by the percentage of the sources using that term as the preferred term (from all sources mentioning the concept). In computing that percentage sources might be weighted. The use of a term as the preferred term in an important source may thus count 5 points while in a marginal source it would count only 1 point. As sources one might also use documents dealing with a concept to see what terms are used in the title, the abstract, or free indexing. An empirical study has shown that in the majority of cases the situation is similar to that of *Humor* (preferred term) and *Wit* (nonpreferred synonymous term). Of 14 articles indexed by *Humor,* 7 contained in their title the term *Humor,* only 1 the term *Wit,* and 6 neither term.

F0.4.3 Criteria for the selection of descriptors

The real problems arise in the selection of the descriptors, that is, the concepts to be included in the indexing language (to be used in document representations and search request formulations). The indexing language, especially the listing of the checklist descriptors, should be displayed in a form that can be grasped easily. (Checklist descriptors are those descriptors that are of special importance in searching and that, therefore, require special consideration in indexing.) Therefore, the number of descriptors in the indexing language, at least the number of checklist descriptors, has to be limited. Often there are additional considerations for limiting the size of the indexing language, especially considerations related to the technical devices used in the ISAR system. On the other hand the advantages of specific indexing, necessitating a larger indexing language, should be carefully weighed.

The selection of descriptors should be based "on their estimated usefulness in communication, indexing, and retrieval". The following criteria are helpful to determine the usefulness of a concept and for making selection decisions:

(a) *Usefulness for searching.* Is the concept likely to be used in search requests? The frequency of occurrence of a concept in previous search requests may

be used as an indicator on this score. Frequency data can be gathered from search requests solicited as a source for the development of the thesaurus, from search requests used in test runs, and from search requests collected during the operation of the ISAR system (for updating the thesaurus). However, one should be aware of the problem discussed in Section B5.3: useful descriptors might be omitted if indexing is too biased towards the search requests presently received or expected. (It has been suggested that "When a word appears relevant for a number of questions, however, it may decrease in value as a search word for an indexing thesaurus because its wide applicability may lead to retrieving irrelevant information." This argument is not valid because high frequency of occurrence in search requests does not necessarily mean high frequency of occurrence in indexing documents.)

(b) *Alternative solutions.* Are there alternative solutions that might be adopted instead of selecting a concept as descriptor? Possible alternatives might be:

(b1) The concept can be expressed by a combination of semantic factors already available as descriptors.

(b2) The concept can be consolidated with a closely related concept, resulting in a newly formed "ISAR concept".

(b3) A broad concept can be used in indexing and searching instead of a specific one.

The availability of alternative solutions is of utmost importance in the selection or rejection of a concept as descriptor. Therefore, meaningful selection decisions cannot be made without taking into account the classificatory structure.

(c) *Logical structure.* Does the concept have "a pertinent relationship with a broader (or narrower, D.S.) subject that was being treated whereby its selection would help to fill out a useful pattern?"

(d) *Frequency of use of a concept in indexing.* This criterion requires more elaborate consideration; therefore it is postponed until Section F0.4.4.

(e) *Number of sources* (thesauri, dictionaries, abstracts, etc.) in which the concept occurs, regardless of what terms are used in these sources to designate the concept. The number of sources in which a concept occurs indicates the importance of that concept. Again, sources might be weighted in computing the frequency. The occurrence of a concept in an important source may thus count 5 points; the occurrence in a marginal source may count only 1 point. In many cases where this criterion is used terms are counted instead of concepts.

(f) *The selection of concepts of general application* (which are often frequently used concepts) requires special considerations (see Section C4.3).

In deciding whether or not a concept should be selected as descriptor the first three criteria (usefulness for searching, alternative solutions, logical structure) are most important. The frequency criteria (d) and (e) are mainly useful in hinting at solutions that then need to be supported by other considerations.

For selecting a concept as checklist descriptor, usefulness for searching is the overriding criterion. It is even useful for this purpose to use a stricter

formulation of this criterion: Is the concept under consideration of importance for the program of research and development (for the planning of the city, for preparing of political moves, etc., depending upon the purpose of the ISAR system)? Is it very likely that it is going to be important?

For the concepts to be included in the indexing language but not as checklist descriptors a less stringent selection is appropriate. If it is not too important to limit the size of the indexing language one should include very specific concepts, too.

On the problem of what compound concepts to introduce as precombined descriptors (rather than using a combination of descriptors) see Section C2.7, especially C2.7.1. The criteria given there partially overlap with the criteria for descriptor selection given here.

Only those concepts that obviously have no relation to the subject area of the thesaurus should be eliminated altogether. The corresponding terms can be left out at the collection stage (Step F1) so that one doesn't need to bother with such cases in the following steps.

In this section we have dealt with the selection of concepts and terms designating the concepts from a collection of terms gathered from different sources. It seems appropriate at this point to emphasize that a somewhat opposite activity is at least as important: the clarification of concepts and the definition and introduction of new concepts and terms to supplement the indexing language and contribute to its logical coherence. We expounded on this point in Sections B7 and C1.1.

F0.4.4 The use of frequency data in the selection of descriptors (technical)

,0 Introduction. Since there is much confusion about the use of frequency data, some clarification is in order. First of all, it must be clear what is being counted, terms or concepts. Second, either of these can be counted from sources like other thesauri, from search requests, or from indexing and/or occurrence in titles, abstracts, or full text documents. (If we want to count terms occurring in titles, abstracts, or full-text documents we have an additional problem: it is easy to identify single words automatically, but it is difficult to identify multiword terms. The same problem occurs in automatic indexing, as discussed in Section B6.2, but there we assumed that multiword terms were already in a thesaurus, whereas here the task is to build the thesaurus. Methods to detect multiword terms are discussed in Chapter H.)

A frequency count on terms is useful in selecting the preferred term from a class of synonyms and quasi-synonyms, as discussed in Section F0.4.2(e). In more sophisticated procedures it can also be used to detect synonyms, as described in Chapter H. For the selection of descriptors we need a frequency count on concepts.

The frequency of occurrence of a concept has to be computed as the

sum of the frequencies of all the terms designating that concept. In many studies this point is overlooked, and term frequencies are used where concept frequencies would be appropriate. (A related and somewhat tricky point is the following: Suppose we have a concept *A* and three narrower concepts *B, C,* and *D*. If *A, B, C, D* are all seldom used, we may not consider them to be good descriptor candidates. However, if we do not use *B, C,* and *D* as descriptors and say "USE BT *A*" instead, we have to sum up all frequencies to obtain the new frequency of *A*. This *new* frequency may then suggest that *A* should in fact be a descriptor, or it may still be so low that we should rather say "USE BT *A'*", *A'* being broader than *A*.)

Concept frequencies from search requests are more important for descriptor selection than concept frequencies from documents. The use of concept frequencies from search requests in descriptor selection is straight forward. Concept frequencies from documents are more difficult to interpret. Since they are usually more readily available, Section F0.4.4,2 is devoted to their interpretation. Section ,1 deals with the collection of data from search requests and from documents.

,1 Gathering of frequency and co-occurrence data. Frequency data can be gathered from:

—search request statements and search request formulations;
—the indexing done during the collection of material, from abstracts, and from full-text documents (these are types of sources mentioned in Section F1.1.2);
—the test run (discussed in Section F6);
—other operating ISAR systems;
—the operation of the ISAR system for which the thesaurus has been built (these frequencies are used for updating).

Another problem is how to actually obtain a frequency and co-occurrence count. This is very easy in mechanized ISAR systems. In manual systems it is difficult. In a card catalog one may check to see whether the volume of cards filed under a descriptor has become too large. (This procedure is facilitated if each descriptor has a guide card with a tab.) Still the catalog has to be scanned regularly. With edge-notched cards or peek-a-boo cards it is difficult to obtain any statistics at all. One possibility is monitoring the frequency of descriptors while searching. (If the search results show that a descriptor is used very frequently or very rarely, one may take action on this particular descriptor.) But this is a haphazard kind of procedure. With peek-a-boo cards descriptors that are used very frequently or very seldom can be selected just by going through and having a short glance at every card. With additional effort it is even possible to get association measures for specific pairs of descriptors. (There is an apparatus that counts holes in Termatrex cards (peek-a-boo) or combinations of those cards.)

The possibilities of data collection in mechanized ISAR systems are illustrated by the plans formerly developed by ASTIA to produce three listings to provide the thesaurus builder with frequency data and related information:

Example:

Descriptor frequency listing.

Descriptor	*Frequency in indexing*	*Frequency in searching*
Jet planes	*2216*	*37*
Jet sea planes	*22*	*9*

Low-frequency descriptor manual file.

Descriptor	*Document numbers*
Alpha chambers	*AD 204 929*
First aid kits	*AD 219 127*
	AD 222 912

This file can be used to assess the value of the infrequent descriptors by looking at the documents. (In addition this file is very useful for retrieval; in searching for infrequent concepts manual look-up is faster than computer search.)

List of context descriptor sets.

Aircraft	*5325*	*(total frequency)*
Co-occurring with		
Engine	*2733*	*(co-occurrence frequency)*
Wing	*2201*	
Rudder	*2182*	
Stabilizer	*2180*	
Airframe	*2023*	
Fusilage	*1845*	
Autopilot	*1673*	
Supersonic	*1580*	
Rotor	*1512*	

Such lists are useful for the more sophisticated methods dealt with in Chapter H.

From some mechanized ISAR systems frequency counts are available (see references in Appendix 2).

The frequency of a concept in an operating ISAR system has to be judged with a view to the following factors:

—relatedness of that system to the system for which the thesaurus is being built;

—size of the collection;

—age and subject field of the collection;

—time elapsed since the first use of the concept within the ISAR system and increase of the collection within that timespan;

—rules used in indexing (if generic posting in indexing is used—i.e., with a specific descriptor all the broader descriptors are to be used in indexing as well—the count of the more general descriptors is inflated);

—frequency of the concept at hand as compared with the frequency of other concepts. If the ISAR system uses very exhaustive indexing, resulting in a large number of descriptors per document, descriptor frequencies in general tend to go up. It might therefore be better to use the rank of a concept in a list arranged by decreasing frequency rather than frequency itself.

Frequency counts (for both terms and concepts from both indexing and search request formulation) can be refined if descriptors are weighted or ranked within document representations or search request formulations. For example, if a descriptor occurs in an important position it is counted 2 or 3 instead of just 1. Or one may simply select a concept as descriptor if it has been used among the four most important terms in indexing any one document.

A quite different method for weighted frequency counts is weighting by source, assigning a higher weight if the term or concept occurs in an important source than if it occurs in a marginal one. This method is particularly appropriate if statistics are based on a count of the number of other thesauri and similar sources in which the term or concept occurs.

Remark: In a situation where documents indexed by free terms serve as sources the following modified procedure for weighting by source has been used: it is possible that the term profile of a document contains only terms that, due to low frequency, would not qualify as descriptors. Thus, none of the terms used to index the document would be included in the indexing language, and the document would not be accessible at all in retrieval. In order to avoid this the weight of a document is decreased each time one of its index terms is selected as a descriptor. (In the beginning all documents have the same weight.) After each weight modification the frequency count is done all over again. This enhances the chance of documents that are indexed only by seldom-used terms to have at least some of their terms included in the indexing language. This may be useful in a fully automated selection procedure but not in the manual or computer-assisted selection procedures recommended in this book.

,2 Use of frequency data in descriptor selection. First of all, frequency data, especially those gathered from other ISAR systems, can give broad hints only. The selection decisions have to be based mainly on substantive considerations. Frequency data from the operation of one's own ISAR system are more useful and should be collected on a continuing basis (if this is

possible without too much effort) as indicators of the need for thesaurus updating. The following considerations hold for initial thesaurus building as well as for thesaurus updating.

Frequency data identify concepts that occur either very frequently or very rarely.

(1) *Concepts used very frequently.* If a concept occurs very frequently in documents, it does not have much discriminatory power in searching if it is used alone. If it is also used very seldom in searching, its usefulness is in doubt. If, however, the concept is used with reasonable frequency in searching, one should investigate to determine which of the following explanations applies:

(1.1) The concept is of general application and mostly used in combination with other concepts. This type of concept can be very useful in searching, as has been discussed in Section C4.3.

(1.2) The concept pertains to a specific subject field and is often used by itself (as the "thematic" concept) in search requests. In this case further subdivision should be considered.

(2) *Concepts used very seldom.* If a concept occurs very seldom in documents, it has very high discriminatory power. If such a concept is used frequently in searching, this high discriminatory power is very welcome. For example, a concept used for indexing seven out of a hundred thousand documents (0.007%) and occurring in 5% of the search requests is of tremendous usefulness in searching and should be considered as a strong descriptor candidate. In fact, this concept is much more useful than a concept used for indexing five thousand documents (5%) and occurring in 1% (or only 0.1%) of the search requests. On the other hand, if the concept is used seldom in searching, it may be too specific, and a USE instruction to a broader concept or to a combination of semantic factors might be appropriate in order to keep the indexing language within reasonable limits. In order to achieve specific indexing it might often be useful to retain as descriptors those low-frequency concepts that belong to the central areas of the thesaurus.

Note: In the case of a concept newly introduced in the subject field no conclusions should be drawn from low frequency.

The above considerations can be formulated more precisely in terms of cost-benefit analysis: the inclusion of a concept in the indexing language incurs costs (larger files, indexing more difficult as size of indexing language increases, etc.). These costs have to be distributed over the documents indexed by that concept. If these documents are few, the cost per document is high. This cost can be justified only if there is a corresponding benefit on the searching side, that is, if the concept in question is used often in search requests.

(3) *Co-occurrence data.* If two concepts co-occur heavily, one should check to determine whether the compound concept formed by their combination should be introduced as a precombined descriptor, using the criteria given in Section C2.7.1.

The considerations of this section partly overlap with Section C2.8.2

on the optimization of an indexing language. More sophisticated uses of frequency data, both for terms and for concepts, in thesaurus-building will be described in Chapter H.

F0.4.5 Central area versus peripheral areas

In selecting the terms to go into a thesaurus, especially the descriptors, one must have a clear picture of the relative importance of the areas to be covered in the thesaurus. One should distinguish:

—central areas;
—areas of intermediate interest;
—peripheral areas.

Many specific descriptors are needed in the central areas; in the peripheral areas a few broad descriptors might do. This difference in emphasis should also be reflected in the lead-in vocabulary, but not as strongly as for the descriptors. An indexer might come across a fairly specific term of a peripheral area and will need to know what descriptor to use.

F0.5 Use of a Thesaurus Form and Related Problems

For the construction of a thesaurus, thesaurus forms on index cards are indispensable (except if very sophisticated automated methods are used). We shall refer to the use of these forms repeatedly as we describe in detail the procedures for thesaurus building. If the necessity of using a thesaurus form is accepted and if the lay-out shown in Figure 54 is deemed useful, Figure 54 may be used as a master form. Instructions for its use are given in Section F0.5.1. The interested reader will find the reasons why a thesaurus-form is needed in Sections F0.5.2–F0.5.3 and the reasoning behind the lay-out and discussion of alternatives in Sections F0.5.4–F0.5.6.

F0.5.1 Instructions on how to use the thesaurus form (technical)

Everything except the top line is self-explanatory. The hierarchical level should be marked by putting "+" (for descriptors) and "−" (for preferred terms) after the appropriate number. If the hierarchical level exceeds 6, the number has to be written in the blank box. Marking the hierarchical level provides a very easy means of giving instructions for typing the classified index from thesaurus forms.

DS (Descriptor), OP (Other preferred term) and CH (Change in existing term) have to be marked, if appropriate, no matter what procedure is followed. DS and OP may be omitted, since the same information is expressed by "+" and "−" after the hierarchical level; however, DS and OP give added protection against errors. Instead of marking NP one may merge

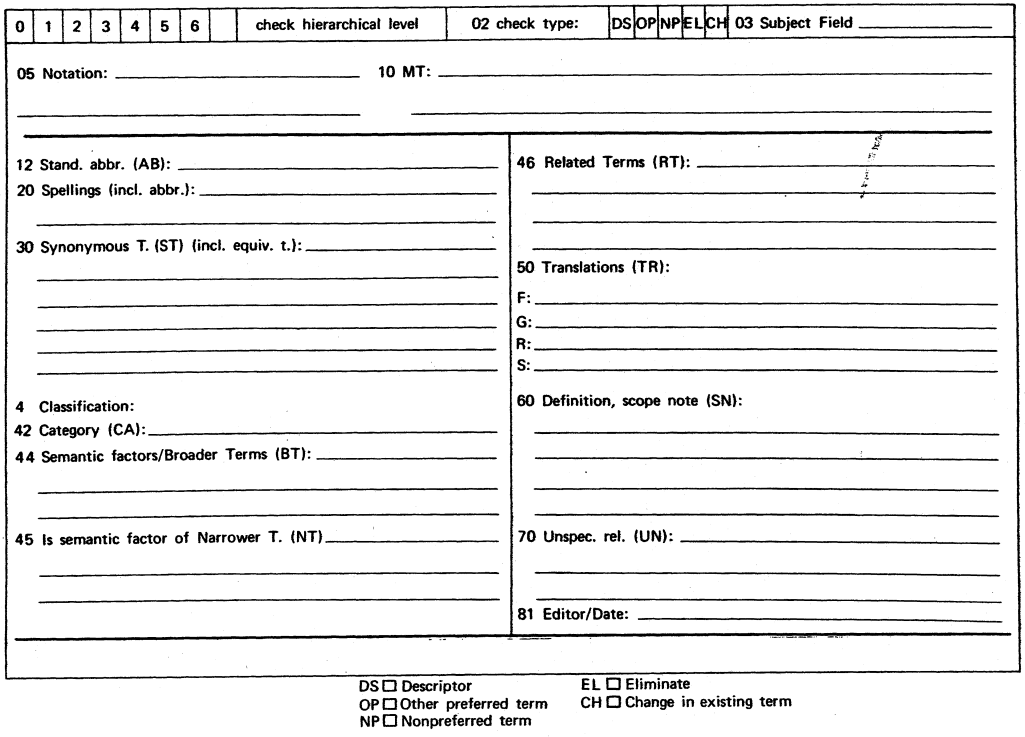


Figure 54. Thesaurus form (F0.5).

the information on the card for the preferred term and discard the card for the nonpreferred term. Instead of marking EL one may simply discard the card. However, even with manual procedures it is easier for the professional just to check EL on the thesaurus form and have it eliminated by a clerical assistant than to eliminate it himself. (Cautious people keep cards to be discarded in a back file until the thesaurus is finished; some keep them even longer.)

As with every form some procedure is needed in case the space allocated for some field is not sufficient. To indicate a continuation use a circled number and put the overflow, identified by the circled number, on the back of the card or on a second card.

For an example of a thesaurus form where the information is filled in see Figure 58 (Section F2.2) and Figure 63 (Section F5.1), where the same form is shown after it has been processed further.

F0.5.2 Reasons for having an index card for each term

In the procedure of constructing a thesaurus it is useful to have an index card for every term so that the terms can be sorted into various arrangements. This holds particularly for the manual performance of Step F2 "Sort into alphabetical order and merge information on identical terms on one card". (If F2 is performed by a computer, no cards are necessary for this step.) Other points where terms have to be sorted and where, therefore, index cards are essential are the steps F3.1, F3.2, and F3.3, where the classificatory structure is worked out.

F0.5.3 Reasons for having a form rather than blank cards

If we had all the information for each term from the beginning, then it would be easiest to use blank cards and to put down the different data elements, properly labeled, one after the other (this procedure is followed, for example, by BASF1). In reality, however, we have a quite different situation. The information to be entered for a term accumulates gradually during the construction of the thesaurus. For example, we may come across Related Terms at different points in our work. So that all these Related Terms can be entered at one place a space has to be reserved for Related Terms. The same holds for other types of cross-references and data elements Therefore, we need a thesaurus form such as the one depicted in Figure 54. However, it would not be practical to have a separate data field for each of the cross-reference types listed in the detailed subdivisions in Figure 21 (Section C7). Accordingly, rather than establish the data fields *Broader Term–Class inclusion* and *Broader Terms–Whole*, for example, and provide a space for each, we just establish one data field *Broader Terms*. If one wishes to pre-

serve the fine subdivisions, this can be achieved by the use of the detailed labels as discussed in Section C7.

F0.5.4 Size

Letter-size (about European size DIN A4) is too large for easy handling. Four by six cards do not provide enough space. Therefore, five by eight (or European size DIN A5) cards are recommended.

If information is entered on the card mainly by handwriting or cut-and-paste techniques, the lines should be parallel to the longer edge of the form ("Querformat") (see Figure 54). This is handy and allows for easy transfer of information by cutting and pasting techniques. A disadvantage is that in typing the main part of the thesaurus one always has to think of the two columns on the form.

If a large amount of information is entered on the forms by typing, the two-column format is definitely awkward. The lines should then be parallel to the shorter edge of the form ("Hochformat"), and the data fields should be arranged sequentially. For the use of a punched paper tape typewriter, to be described in Section F9.1.1, "Hochformat" is mandatory. However, the disadvantages are that it is less handy and file drawers are not as easily available.

F0.5.5 Width of lines

The form depicted in Figure 54 uses lines corresponding to 1½ line spaces on a typewriter. This is convenient for filling in by hand. If the forms are to be filled in by typewriter, tabulator, computer printer, or other equipment, the lines have to be adjusted accordingly.

F0.5.6 Sequence of data fields

The data fields should be arranged on the thesaurus form in the sequence to be used in the user version of the thesaurus to simplify transferring the data in the production of the user version. The sequence of fields in the form depicted in Figure 54 agrees with the discussion of sequence of fields in Section D2.2.

F0.6 Working File and User Version

We have repeatedly referred to the working file and to the user version of the thesaurus. The working file is used in working out the thesaurus and updating it. It contains the most detailed information. The user version of the thesaurus, to be used by indexers and searchers, need not give that much detail, e.g., with respect to source indications or the distinction of cross-reference types. The physical form of the working file should be such that modi-

fications and additions can be made easily; that means either a card file on thesaurus forms or a computer-stored file. The physical form of the user version should be such that it can be easily used and that it can be reproduced at reasonable cost; that usually means book form.

More precisely, the working file corresponds to the main part of the user version. The classified index and the alphabetical index can be produced from the working file also. The working file is thus the master file of the thesaurus from which all parts of the thesaurus can be produced. Especially with computerized procedures one might consider storing only the working file and produce the classified index and the alphabetical index from the updated working file each time a revised edition of the user version of the thesaurus is prepared. This would simplify storing and updating the thesaurus. The alternative would be to maintain working copies of the classified index and the alphabetical index and insert revisions as they arise. It depends on the circumstances which of the two solutions is cheaper, but we suspect that usually it is more efficient to store and update the classified index and the alphabetical index separately. Note also that a working copy of the classified index reflecting the latest changes is very helpful in processing further revisions.

F0.7 Source Indications for Data Elements Entered in the Thesaurus

F0.7.1 Why source indications?

,1 Use of the source indications for the elaboration of the thesaurus. Source indications are useful for the elaboration of the thesaurus. It is, for example, possible to look up the place of a concept in the hierarchy of a classification scheme used as a source. This might give suggestions for the building of one's own hierarchy. One could look up the definition of a term, or one could check to see how a term is used in its context in the abstract that has been used as a source. Also it might be useful to look up the frequency of a terms in the ISAR system that employs a particular classification scheme or thesaurus used as a source.

,2 Why source indications in the user version of the thesaurus?

(1) *Reference to definitions.* Some sources contain a definition and/or further explanations of the preferred term; the user should be referred to such sources by an appropriate source indication given as part of the scope note (data field SN), as described in Section C3.2.1. (In the working file these sources appear also in the specific data field, as appropriate; see below.)

(2) Source indications in thesauri are especially important in the context of *cooperation in information services.* For example, through source indications it is possible to determine the institutions that use a particular descriptor (and are therefore likely to have material on that descriptor) and

possibly the form in which this descriptor is used in each system (cumulative thesaurus). Such sources should be given in data field SR following the scope note or, if the exact form of the descriptor as used in the source is given in a cumulative thesaurus, in the format described in K1.3.3.

F0.7.2 Keeping track of the sources in the working file (technical)

(For examples, see Figures 56–58, Section F2.2, and Figure 61, Section F2.3; it might facilitate understanding to skip Section F0.7.2 now and come back to it after reading F2.)

Recording and keeping track of the source indications is a somewhat tricky matter. First, one has to decide how detailed the source indications should be. In the detailed form the exact sources for each and every data element entered in the thesaurus are kept precisely. In a computerized procedure this is easy. However, with manual procedures the effort may be prohibitive and not worth the benefits. In this case one should use the crude form. In the crude form one only keeps track of the sources in which a concept as such occurs and what term is used in each source to designate the concept. One does not keep track where ST, BT, NT, and RT cross-references and other information on the concept come from. (If the need arises, one may check each of the sources mentioning the concept to find out the source from which a certain data element came.)

Keeping track of the sources comes in at two points in the procedure for thesaurus building:

(1) *Transfer from sources.* In the process of transferring a term and information on that term from a source to a thesaurus form it is sufficient to give a source indication after the Main Term in data field MT. It is understood that all other data elements on the card come from the same source. If the detailed form is to be used in keeping track of the sources, the source code should be underlined in this step for reasons that will become clear shortly. The following procedure is designed for the case where only entries for terms that are preferred terms in the source are transferred on thesaurus forms (this is in line with Section F1.2.1,1, "Preparation of pre-arranged sources").

The format for the source indication is as follows:

(Source code: Notation from source/frequency given in source in percent).

Notation and frequency are simply omitted if they do not appear in the source.

(2) *Merging information from different cards.* The second point is when information from several cards (thesaurus forms) is merged on one card. For simplicity we assume that there are only two cards, card 1 and card 2, and that the information is to be merged on one card. We deal with the crude form first and then proceed to the more complicated procedure needed for the detailed form.

(2a) *Crude form.*

(2a1) The Main Term on card 2 is the same as the Main Term on card 1: Enter the source indication from card 2 in data field MT of card 1. Simply transfer

all other data elements from card 2 to card 1 without paying attention to source indications.

(2a2) The Main Term on card 2 is a synonym of the Main Term on card 1: Enter the Main Term from card 2, together with its source indication, in data field ST of card 1. If the term is already given in field ST of card 1, just add the source indication. Transfer all other data elements as in case (2a1).

Spelling variants can be dealt with in the same way as synonyms. However, if spelling variants are not important in the thesaurus to be built, one might disregard differences in spelling.

Card 2, to be merged on card 1, might in turn be the result of an earlier merge. In this case terms from the data fields MT and ST are transferred together with any source indications that might already be attached to them.

(2b) *Detailed form*

Now the source for every single data element is kept. Therefore, before transferring any other data elements to card 1, the source code given in field MT of card 1 is added to every data element in every data field of card 1; the source code is *not* underlined.

(In most situations it is sufficient to give just the source code, not a full source indication with notation and frequency for BT, NT, and RT. In some situations it might be useful to have the notation, too, for BT, NT, and RT, at least for selected sources. In this case a source indication for each term entered in these fields has to be made while transferring entries from the source to thesaurus forms.)

The further procedure is as follows:

(2b1) The Main Term on card 2 is the same as the Main Term on card 1: The source indication from data field MT of card 2 is entered in field MT of card 2 (the source code underline is carried!). Further information from card 2 is transferred as follows (using data field RT as an example): If card 2 gives an RT already on card 1, only the source code from data field MT of card 2 is added. If card 2 gives an additional RT, that term is entered in field RT of card 1 with the source code from MT of card 2 (the underline is *not* carried in this case). In this way no confusion about the sources of a data element can occur.

(2b2) The Main Term on card 2 is a synonym of the Main Term on card 1. The Main Term from card 2 is entered in data field ST of card 1, together with its source indication. If the term is already contained in data field ST of card 1, merely the source indication is added. In either case the underline under the source code is carried to show that the term is the preferred term in the source. Further information is transferred as in case (2b1).

A special problem can occur with synonyms, as illustrated by the following.

Example:

Card 1: *Lawyer (CT)*
ST Attorney (CT)

Card 2: Attorney (WH)
ST Lawyer (WH)
After the merge, card 1 looks as follows:
Lawyer (CT) (WH)
ST Attorney (CT) (WH)

Card 1 or card 2 or both might be the result of a previous merge. In this case all source indications are transferred (underlines under source codes are carried).

In the description of the procedure for thesaurus building in Sections F1 ff. we shall repeat parts of this section at the appropriate points.

Special considerations on the source indications are necessary if one wants to build a cumulative thesaurus as described in Section K1.3.1.

F0.7.3 Experts and lexicographers as sources (technical)

Input into the thesaurus comes not only from other thesauri and from documents but also from consulting scientists, panel discussions, and decisions by the editor(s)/lexicographer(s). As far as practical, these sources should, for internal purposes, be treated as all other sources. Keeping track is especially difficult in these situations, however, and may not be worth the effort required.

Part of the difficulty arises from the fact that in the very important step of hierarchy building one does not deal with each term (or the card for each term) individually but with whole groups of terms that are rearranged continuously until a satisfactory arrangement is found.

F0.7.4 Keeping track of deletions (technical)

The procedures outlined above do not provide for the possibility of recording decisions on initial rejection or on deletion of data elements. If one wants to keep track of those decisions the easiest way to do so would be as follows: The data elements initially rejected or to be deleted are kept in the working file but tagged by an appropriate symbol. In typing the thesaurus (or in printing it out by a computer) these data elements are then omitted. The procedure described in Section F0.8 incorporates this feature.

F0.8 Keeping Track of Decisions and Dates

In the working file it might be useful to record who made a particular final decision on the data element (that includes decisions on the inclusion of a descriptor). In both the working file and the user version of the thesaurus it might be useful to keep the dates when a certain data element has been entered. It might also be useful to treat deletions in the same way. In many cases, however, the effort to do so is not justified; one should carefully weigh costs against benefits. The procedure described in the following makes it

possible to keep track of every minute detail. A less detailed procedure might be appropriate in many circumstances. Also, the procedure is described in terms most appropriate for computer applications. The principles are the same in manual application, however.

F0.8.1 Keeping track of decisions and dates in the working file (technical)

,1 Keeping track of decisions made. Whenever a decision on a record as a whole is made a fixed-field decision indicator string is entered as the active string in data field 81 "Editor and date when entered" (Figure 21, Section C7); the former active string (if any) becomes inactive. An example of a decision indicator string is:

1 ETS 67-07 68-12 $

The elements of the decision indicator string are as follows:

—Status code;
—Initials or other code of the editor/lexicographer making the decision;
—Date when record entered into the working file;
—Date when record entered into the user version;
—End mark.

The status code is as follows:

0 to be entered into user version
1 entered in user version
4 to be deleted from user version
5 deleted from user version

(In the construction phase only, a simpler procedure can be used for the elimination of whole entries: put EL in data field 02 Type.)

For even status codes the second date is blank. When the record is printed in or deleted from the user version, the appropriate date is entered. For all records in the first edition of the thesaurus, this is the date at which the first edition of the user version has been completed. The information contained in data field 81 for the record as a whole can be given for a single data element. The appropriate decision indicator string(s) are enclosed in brackets (or other delimiters) and follow the source indications (if any). The active decision indicator string for a single data element overrides the active decision indicator string for the record as a whole, except if the whole record is to be deleted.

Things get just a little more complicated if one has to keep track of the inclusion of a change into a supplement, a cumulative supplement, and finally a new edition of the user version.

The procedure described takes care of changes from descriptor to non-descriptor, and vice versa, due to the fact that, for example, "term is descriptor" is expressed by the data element DS in data field 02 Type. Note also that

EL (Eliminate) in data field 02 is allowed during the construction of the thesaurus only.

In a cumulative thesaurus the situation is more intricate, as discussed in Section K1.3.1.

‚2 Keeping track of decisions still to be made. The data fields 82–86 provide the possibility of recording where decisions have been postponed and where necessary information may be obtained.

An X is put into data field 82 (or a paperclip on the thesaurus form) if a record is not yet final and a decision cannot be made right away. It is then possible to single out at any time those records that need further work.

The data fields 85 "Name of expert to be consulted" and 86 "Question to be asked" make the consultation of experts more efficient. A duplicate of each pair 85/86 is kept in a file sorted by experts. Experts and questions can be written on query-and-reply slips. From time to time the questions can then be asked, either orally or in writing. This procedure has the advantage that all the questions to be asked of any one expert are batched.

‚3 Keeping track of why decisions have been made. For later reference it is useful, at least in some cases, to note down the reasons why a particular decision has been made. This type of "documentation" can take several forms. One might write an essay giving the rationale for the over-all arrangement. (It might even be useful to include such an essay in the introduction to the thesaurus.) Considerations on the subdivision of a whole subject field can be given in the scope note (data field 60 SN) if they are useful for the thesaurus-user, or in the internal scope note (data field 61 SN-IN, cf. Figure 21, Section C7) otherwise. The same holds for comments on individual descriptors.

Keeping track of the reasons for decisions is especially difficult in meetings in which the thesaurus is discussed. If it is not possible to enter a summary of the discussion on the thesaurus form during the meeting, one has to keep minutes and transfer the information to thesaurus forms later on.

F0.8.2 Giving dates in the user version of the thesaurus (technical)

Some dates are of interest to the user of the thesaurus: the date when a descriptor has been actually included in the thesaurus for use in indexing, or when a descriptor has been deleted, and possibly dates when some of the cross-references have been introduced. These dates are best given in the scope note for the descriptor. In the working file these dates are stored with the appropriate data element, as described above, in addition to their appearance in the scope note.

After the discussion of these general problems we can now go on to describe the individual steps needed in the construction of a thesaurus. Some

of the descriptions are rather technical. The reader might find it useful actually to work out an example in order to gain a better understanding.

F1 COLLECT AND RECORD MATERIAL (CONCEPTS, TERMS, RELATIONSHIPS BETWEEN AND AMONG THEM)

It is natural and useful to start the development of a thesaurus by gathering, from a variety of sources, information as complete as possible on concepts, terms, and all kinds of relationships between terms and concepts (synonym-homonym structure and equivalence structure) and among concepts (classificatory structure). Based on the material so collected, one can then develop the structure of the thesaurus and introduce necessary additions.

F1.1 Kinds of Sources. Criteria for Selection of Sources

F1.1.1 Sources in which terms are already arranged according to some principle (prearranged sources)

(a) Descriptor lists, classification schemes, thesauri (this includes universal classification schemes such as LCC or UDC, or parts thereof, and special classification schemes, e.g., schemes used in special libraries, patent classification schemes);

(b) Nomenclatures of single disciplines such as the nomenclature approved by IUPAC (International Union of Pure and Applied Chemistry);

(c) Treatises on the terminology of a subject field or subfield;

(d) Encyclopedias, lexica, dictionaries, glossaries (universal or discipline-oriented; mono-, bi-, or multilingual);

(e) The tables of contents and indexes of textbooks and handbooks;

(f) Indexes of journals and abstracting journals;

(g) Indexes of other publications in the field;

(h) Term-association lists produced by subjects in term association studies or similar experiments (see Section F1.1.4).

(i)–(j) (Reserved for additions).

Institutions and bibliographies that can be consulted to find prearranged sources are given in Appendix 2.

F1.1.2 Sources in which terms are not ordered or from which terms must first be derived (open-ended sources)

(k) Lists of search requests and interest profiles. Search requests can sometimes be obtained from records of operating ISAR systems. Another approach, to be used instead or in parallel, is to solicit search requests from potential users. It is also possible to have the same users select terms useful for the expression of their search requests.

(l) For ISAR systems in specific institutions: descriptions of the projects in research and development or of other activities to be supported by the ISAR system.

(m) Discussions with specialists in order to identify their interests and potential search requests. In personal interaction one might get a better idea of user needs and points of emphasis than in written answers. The result of such a discussion is a list of terms and themes recorded by the thesaurus builder.

(n) An extremely useful variant of this is the following method: A sample of about one hundred documents representing the scope of the thesaurus to be developed is selected in cooperation with a subject expert. A meeting of seven to twelve potential users is organized. The documents are presented to potential users and for each document one asks the question: What are the aspects under which this document may be of interest to your work? The sample documents serve as stimuli to elicit the explicit formulation of interests that otherwise may have remained hidden. This method yields a large number of concepts and terms that are of immediate interest to the users of the ISAR system. It might be possible to achieve similar results by sending out documents to specialists and asking for written answers.

(o) Have a number of documents indexed by experts in the field or (less desirable) by indexers in the information center or other staff using terms of their own choice; in order that many synonymous terms be collected, it is recommended that the same documents be indexed by different experts.

(p) Titles of documents.

(q) Abstracts and reviews of documents.

Conference programs provide a timely source for both titles and abstracts.

(r)-(y) (Reserved for additions).

(z) Finally, the editor(s)/lexicographer(s)) working on the thesaurus give their own input and should therefore be considered as a source.

Since the indexing language or thesaurus should tell the indexer what aspects are important for the users of the ISAR system and should therefore be considered in indexing (request-oriented indexing as implemented through the checklist technique), the study of user needs provides an input of paramount importance for thesaurus building. If general studies of user needs are available, they should be consulted. Specifically, thesaurus-directed data on user needs are contained in search requests, more generally in the sources (k)–(n). These sources should receive the greatest weight in thesaurus construction. Very often this point is neglected and thesaurus construction is mainly, if not exclusively, document-oriented. This can be justified only if it can be shown for the ISAR system in which the thesaurus is to be used that terms derived from documents are the same as those derived from search requests and that the term frequencies and other indicators of term importance are also the same.

F1.1.3 Selection of the sources to be used

The number of sources to be selected—the completeness of the coverage—is a function of the resources available. In any case one should aim to make the

collection of concepts and terms as complete as possible within the scope of the thesaurus. As will be shown below it is usually not possible to achieve this end by using prearranged sources only or open-ended sources only. Each type complements the other.

The two kinds of sources have the following characteristics:

(1) The *prearranged sources* require less effort in the gathering of material. Often the terms are already in a standardized form. Furthermore, these sources indicate relationships between terms and concepts and relationships among concepts in an explicit way. On the other hand, prearranged sources suffer from the following disadvantages: The viewpoints used in selecting and arranging terms and concepts are often very specific and narrow and/or do not take into account the complexity of the subject field. In most cases too few of the synonyms and quasi-synonyms are given (unless a good thesaurus in the subject field is available already).

One may rely mainly (but not only) on prearranged sources if the following conditions hold for the subject field in which the thesaurus is to be developed:

—recognized special classifications and thesauri, extensive and extensively cross-referenced indexes of abstract journals, and larger terminological works are available;

—nomenclatures for materials, living organisms, etc., are available;

—the field is not in a phase of rapid development.

(2) *The open-ended sources* require more effort in information gathering. They have the advantage of yielding a complete collection of those concepts that are necessary to express the subjects asked for in search requests. These concepts are identified in the degree of specification in which they occur in the search requests and in the documents. The terminology reflects the actual usage in the field. Furthermore, the collection reflects the current conceptual and terminological status of the field, not the status of five or fifty years ago. Therefore, these sources should be specially emphasized for mission-oriented thesauri, for thesauri in complex subject fields, and for new, highly specialized or fast-developing subject fields.

In selecting the sources one should make sure that the whole area of the thesaurus is covered. In using prearranged sources one should be careful not to neglect marginal areas.

Furthermore. the following criteria can be used for the selection of prearranged sources:

"1. They contain scientific and technical terminology. (With other thesauri the appropriate field has to be substituted here, of course. D.S.).

2. Their development was from the actual indexing (and searching, D.S.) experiences, thereby representative of storage and retrieval requirements.

3. They were strong in thesaurus-like arrangement, showing various kinds of cross-referencing data, generic relationships, scope notes, and frequencies of use."

It is important to select a representative sample of open-ended sources. With search requests or user discussions this might be difficult. With docu-

ments or abstracts it is easier. A reasonable sample size might be 1,000–2,000 abstracts. The sample may be obtained by scanning relevant journals and/or abstracting journals and/or by asking potential users to submit relevant documents.

Remark: In selecting documents to serve as sources of terms (be it from the table of contents, from an abstract, or from free indexing terms), one should take care to include both pre-research documents (proposals, descriptions of research projects) and post-research documents. It has been observed in a study in the field of neurological diseases that "it is apparent that the semantemes of high frequency in the pre-research documents and of low frequency in the post-research articles are rather general terms, while those that are of high frequency in the post-research articles but low in the pre-research documents are specific and tend to be clinically oriented." It might be possible to detect hierarchical relationships by comparing the terms used in pre- and post-research documents on the same research project.

F1.1.4 Term-association lists (special topic)

Term-association lists obtained from subjects representative of the user group are an especially useful source since they reflect the conceptual and terminological "map" of the user. Term-association lists are on the borderline between prearranged sources and open-ended sources.

There are two methods of obtaining term association lists. We might call them the *free association method* and the *bound association method.*

In the free association method each individual is presented with a number of terms, the stimulus terms, and asked to name any terms that he thinks of in connection with each stimulus term. In this method new terms are added to the initial vocabulary. In studies done with this method terms in both definitional and contextual contiguity relationship to the stimulus term are named by the subjects.

The bound association method can also be described as a brute-force approach to determine relationships between terms, once the list of terms has been established. Each ordered term-pair is presented to several (for example, 3) subjects, and the subjects are asked to determine the relationship that holds between the two terms in the pair. This procedure is, from a theoretical point of view, very much in line with our considerations on hierarchy-building in C1.2; however, it is impractical in most situations.

F1.2 Technical Procedures for the Recording of Terms, Etc.

F1.2.0 Introduction

For each term to be entered from a source into the initial collection of terms a record has to be established. This record contains the term itself as Main

Term and possibly other data elements, such as Broader, Narrower, and Related Terms. With manual procedures, each term and the data elements for it should be transferred to a separate index card (thesaurus form, see Figure 54, Section F0.5) so that the terms can be sorted easily. Each record thus consists of one card. Having in mind the reader who is interested mainly in manual procedures for thesaurus construction we shall generally use the term "card" instead of the more general "record" throughout Chapter F. (In Chapter D and elsewhere the term "entry" is used with the same meaning.)

Before information on terms can be transferred from sources to cards, the sources have to be prepared as described in Section F1.2.1. The actual transfer of information will be discussed in Section F1.2.2.

F1.2.1 Preparation of sources (technical)

,0 Source identification codes. Each source is identified by a short code which later serves as an indication of origin for all information taken from that source. Any system for the assignment of these codes will do. For prearranged sources the following are examples:

Examples:

(1) A combination of four letters, namely, the first three letters of the name of the author and the first letter of a word of the title. For example, Crad = Craig, R.: The Dynamics of Stratospheric Circulations.

(2) A combination of three letters arbitrarily selected among the beginning letters of authors and/or words in the title, for example, CDS.

(3) Two capitals drawn from the name of the issuing organization, e.g., BY — Boeing Company.

For open-ended sources (search requests, abstracts, etc.), sequential numbers or, if available, call numbers may be best. For scientists as sources and lexicographers as sources initials might serve as source codes as long as they are unambiguous. For purposes of machine processing it is convenient if all source codes have the same fixed length.

,1 Preparation of prearranged sources. From the prearranged sources, terms can be transferred to cards without prior scanning and selection. Prior scanning and selection is recommended only if 25% or more of the terms are likely to be eliminated right away. Otherwise more work is needed for scanning and selection then is saved by eliminating the work of transferring unwanted terms. One should keep in mind that in this phase of thesaurus development only those terms are to be eliminated that are obviously beyond the scope of the thesaurus. If cards have to be made for selected terms only, those terms have to be marked in the source—for example, by "*" or "√".

If the source in question contains USE instructions and if for every USE instruction the corresponding inverse UF statement is given, it is not

necessary to prepare cards for nonpreferred terms. In fact these cards would only create work without adding any new information.

Example:

Television camera tubes
UF Pick-up tubes
Pick-up tubes
USE Television camera tubes.

Pick-up tubes is a synonym of *Television camera tubes.* Therefore, there is no need to make an extra card for *Pick-up tubes;* this card would only be eliminated later on when all cards referring to the same concept are merged (in Step F2.3, "Second round of merging", or in Step F3.3, "Work out detailed thesaurus structure. Select preferred terms"). However, if a USE instruction is of the USE BT type (whether or not it is explicitly so designated), one may want to have a separate card for the specific concept from the beginning.

Example:

Television camera tubes
UF Iconoscopes
Iconoscopes
USE Television camera tubes.

Iconoscopes are a special type of *Television camera tubes,* and a separate card should therefore be established. If this is not done in the transfer operation (where it is merely a clerical process), it has to be done later, while working on the card for *Television camera tubes.* If this case occurs often, one may include initially all terms that have a USE instruction. The cards for truly nonpreferred terms like *Pick-up tubes* are then eliminated in later editing.

Some of the prearranged sources, e.g., TEST and often special dictionaries, are usually too big to be included or even to be searched through for relevant terms. They may, however, be used to look up information on terms obtained elsewhere, as described in Section F2.2.1. Or the terms from certain sections are included (e.g., the terms listed under the appropriate subject categories in TEST).

,1.1 Adding an auxiliary notation. Some sources arrange terms in classified order but do not attach a notation to them. In this case an auxiliary notation is added, using the modified decimal notation described in Section D4.3.4 (the notation is used at a later stage if one wants to refer back to the source, e.g., in step F3.3.2, "Work out the classificatory structure").

,2 Preparation of open-ended sources: mark terms to be transferred. With open-ended sources it is necessary to identify the significant terms before

they can be transferred to cards. In most cases positive selection will be used: all significant terms occurring in search requests, abstracts, etc., are underlined. Even wrong, inexact, or popular terms are to be marked. The same is true for terms that belong to subject fields that are marginal for the thesaurus. Index terms that describe the content of the search requests or the document more precisely and/or on a higher level of abstraction may be added as deemed necessary by the editor. In working with full documents as sources it might be useful to use index terms only.

Terms that occur several times in the same document are taken over only once for this document (possibly recording the frequency within the document). However, if the same term occurs in several documents, several index cards are made up accordingly (compare Section F1.2.3).

Examples:

THE LOW-INCOME FARMER IN A CHANGING SOCIETY

To identify some major differences among low-income farmers, and to delineate the group that represents the real core of the persistently poor, data were obtained from 189 farm operators representing a stratified random sample in Fayette County, Pennsylvania, in 1957. The five main categories of individuals identified were: (1) the aged, (2) the physically handicapped, (3) the farm operator primarily oriented to non-farm opportunities, (4) the farm operator oriented to commercial agriculture, and (5) the farm operator oriented to subsistence agriculture. The characteristics of the core of low income subsistence farmers who normally do not respond to either welfare or economic-development efforts were examined in greater detail. It was found that they: (1) retained traditional values while having lost many traditional subsistence skills, (2) failed to respond to greater agricultural efficiency and productivity efforts because commercial success was not highly valued, (3) placed extreme emphasis on neighborliness and friendliness as their primary goals, and (4) must respond to an attempt to change prestige orientation if their cycle of poverty is to be broken.

NEMATODE CONTROL IN SWEET POTATOES

The yield and quality of sweet potatoes can be increased by soil fumigation or the addition of solid nematocides in some areas of Mississippi. The commercial fumigants Vorlex, Dow W-85, and DD significantly increased yields and quality in the treatments of rows. Vorlex or Dow-85 should be applied at 2.5 gal/acre and DD at 9-10 gal/acre, 8-10 inches deep in the center of the row, 14-30 days prior to planting. Broadcast fumigation was also effective, but required higher fumigant levels. Among the experimental solid nematocides, Bayer 68138 and Dasanit showed promise. This study of control of rootknot nematodes was conducted by the Truck Crops Branch Experiment Station in 1967 on three- and four-row replicated and random-

ized field plots known to be infested with the nematodes. More information is deemed necessary than was obtained from this one-season field test.
Added terms: Application dose; Application time

,3 Pre-processing of open-ended sources. In working with open-ended sources one may also use negative selection, that is, include all terms that are not on a stop-list. This is sensible only if computer assistance is used to produce a list of all the terms occurring in a corpus of open-ended sources.

An intermediary solution in which the open-ended sources are pre-processed is also possible: a listing of all non-stop-list terms occurring in the open-ended sources is produced. From this listing terms to be included in the thesaurus are then selected, possibly using frequency criteria as discussed in Section F0.4.4. Such a listing is particularly useful if the context of each term is given. This might simply be done by producing a KWIC index (the units being titles, search request statements, or sentences from documents and/or abstracts). Such a listing is very useful for the study of homonyms and for the study of relationships between terms and for the formulation of definitions. Further elaboration of these methods leads to the automatic construction of indexing languages, to be discussed in Chapter H.

F1.2.2 Transfer of terms to cards (thesaurus forms) (technical)

After these preparations the terms and other information can be transferred to cards (thesaurus forms), as shown in Figure 55.

,1 Entering Synonymous, Broader, Narrower, and Related Terms. Together with a term, additional information, such as Synonymous or Equivalent Terms, Broader and Narrower Terms (one level up or down), and Related Terms (possibly including Coordinate Terms, i.e., brothers in a hierarchy), short definitions, etc., is transferred to the appropriate data fields of the thesaurus form. (Data fields for which no information is given in the source are simply left blank. Note that data field 05 Notation is left blank for later use; the notation from the source, if any, is given in the source indication; see section ,2.) Long definitions are only referred to. In certain situations, Broader and Narrower Terms need not be transferred from sources that have a classified arrangement, as explained in the rest of this paragraph. If the terms in a source are arranged in classified order, Broader and Narrower Terms for a term given can be easily looked up in the source, using the notation of the term (if necessary, the auxiliary notation assigned in F1.2.1,1). If keeping track of sources of relationships is not an important problem, one may therefore omit the Broader and Narrower Terms from these sources. At the stage described in F3 one consults the original source and applies the information for hierarchy-building. In those cases where the classified arrangement chosen coincides with the classified arrangement in

0	1	2	3	4	5	6		check hierarchical level	02 check type:	DS	OP	NP	EL	CH	03 Subject Field

05 Notation: _____ 10 MT: Color TV picture tube (TH: 659.5)

12 Stand. abbr. (AB): _____

20 Spellings (incl. abbr.): _____

30 Synonymous T. (ST) (incl. equiv. t.): Color kinescope

4 Classification:

42 Category (CA): _____

44 Semantic factors/Broader Terms (BT): _____
Color TV receiver

45 Is semantic factor of Narrower T. (NT) _____
Color TV screen

46 Related Terms (RT): Flat picture tube;
Radiation hazard

50 Translations (TR):

F: _____

G: _____

R: _____

S: _____

60 Definition, scope note (SN):

70 Unspec. rel. (UN): _____

81 Editor/Date: _____

Figure 55. Example of filled-in thesaurus form (F1.2.2).

the source, Broader or Narrower Terms need not be transferred even at this stage. In other cases Broader and Narrower Terms are transferred as is deemed useful. This procedure saves much work, both in the transfer of terms and in merging information from different cards. If, in Section F3, "Work out the structure of the thesaurus", machine processing is to be used, this procedure is not applicable. Compare Section F4.0 to the problem of transferring Broader and Narrower Terms.

If a source uses the crude lead-in method, that is, does not distinguish between UF ST and UF NT, we have the problem of where to enter the terms listed in the source under UF. One may choose between three strategies:

(1) Assume that most UF statements do in fact refer to Synonymous Terms and enter all terms from UF in field ST. Corrections will then be made in later editing.

(2) Enter all terms from UF in field UN (Unspecified relationship). Further specification is then made in later editing.

(3) Exercise judgment during transfer and put terms from UF into SP (Spelling variants), ST, or NT (or sometimes RT), as the case may be. Keeping track of the source precisely presents a problem in this case.

Relationships among terms can also be detected from open-ended sources, such as search requests/interest profiles and abstracts, and should be transferred to the thesaurus forms.

Example for the case of abstracts:

From the second sample abstract given above it can be seen that Dasanit has a broader concept (Experimental) solid nematocides. Therefore, on the card for Dasanit one should enter (Experimental) solid nematocides as a Broader Term.

Search requests are also very useful for detecting relationships, especially if they have been formulated for an ISAR system using natural language as indexing language. In this case the searcher should name as many synonyms designating a certain concept as he can think of and combine them all by OR. In the case of an inclusive search he has to add terms for narrower concepts, too. Looking at search request formulations one should therefore analyze the terms co-occurring in an OR parenthesis to see whether there are relationships of synonymity or Narrower Term-Broader Term relationships or whether a suitable Broader Term, covering all the terms combined by OR, should be introduced. Interest profiles that have been improved through feedback over a period of time are especially useful as a source for this procedure.

These sources can be exploited further by detecting term relationships through statistical methods, as discussed in Chapter H.

,2 Entering the source indication. After the term in data field MT, the source indication is given in the following format:

(Source code: Notation in source/Frequency given in source in percent)

If notation and frequency are not given in the source, they are simply omitted. (For the detailed form of keeping track of the sources only: The source code is underlined. In some cases it might be useful to give the notation for BT, NT, and RT, too. In this case a source indication, omitting frequency, is entered after each term in these data fields.)

In some cases it might be useful to add a page number to the source indication so that it is easier to find the term in the source. This is useful in the construction of the thesaurus and mandatory if it is planned to include a reference to the source in the user version of the thesaurus.

(A more detailed referencing procedure is possible but not recommended: number the entries on each page of the thesaurus and give page and entry number, together with the source indication. This procedure is not recommended, however, because the minor benefits (if any) for the later steps do not justify the major costs in the step of collection of material. Based on a notation or the alphabetical sequence, any term may be looked up rapidly in any source without having a page number—certainly without having an entry number on the page.)

Often it may save labor to stamp the source codes on the cards (e.g., using a rubber stamp printing set). In this procedure the card decks resulting from different sources are kept separate until the source code is stamped on. However, if notations or page numbers have to be added, this method is less practical. In mechanized methods the inclusion of the source codes is even easier.

,3 Transfer of terms and other information with manual procedures. Terms from the open-ended sources have to be typed or written on the cards.

For transfer of the terms from prearranged sources, two procedures are possible:

(a) type or write on cards;

(b) copy the source, cut the entries, and paste on cards. Which of these procedures is cheaper has to be decided from case to case. The following parameters have to be considered in the decision:

—how much text has to be transferred to the index cards? (text may include a definition or scope note);

—machines available (a machine may considerably speed up the pasting of entries on cards);

—clerical staff available (pasting requires less skill than typing and is therefore cheaper!).

If cutting and pasting is used, it is often not possible to fill in the in-

formation in the proper spaces of the thesaurus form. In this phase of the thesaurus development, this is of minor importance, provided that the different data fields (such as Synonymous Terms, Related Terms) can be identified without difficulty.

F1.2.3 An alternative procedure

With the method of term collection suggested here multiple cards are made for a term occurring in several sources, and duplicates are not removed until the next step. An alternative procedure would be as follows: Make cards for the terms of the first source and alphabetize. In processing the next source look for each term in the alphabet. If the term is found, add information to the card. If it is not found, make a new card and insert into alphabet.

It is hard to say whether this method is cheaper. This depends on the number of identical terms and the arrangement of terms in the source: If the arrangement in the source is alphabetical, the look-up procedure may be cheaper; if the arrangement is hierarchical, it is cheaper first to collect and then to eliminate duplicates. An intermediate strategy is also possible: Start with the open-ended sources and with the sources that have a classified arrangement. Transfer terms to cards, as described previously. Alphabetize and eliminate duplicates. Then process further sources that are arranged alphabetically by checking and merging information on the same term and interfiling cards for new terms.

Compare Section F2.2.3 on "pulling" information from a big thesaurus.

F2 SORT INTO ALPHABETICAL ORDER AND MERGE INFORMATION ON IDENTICAL TERMS ON ONE CARD

F2.1 Sort into Alphabetical Order. Rules for Preliminary Alphabetical Sorting

Common sense alphabetical sorting can be used in this phase; consideration of complex filing rules is usually not necessary. The cards for identical or nearly identical (singular/plural or similar variations) terms are put together with a paper clip. Often terms consist of a string of terms separated by commas (for example, the term *Beer, ale, malt liquor*). In this case all terms that start with the same term are considered synonymous and these cards are clipped together. However, this is not always useful, e.g., *Roughness, smoothness* is broader than *Roughness,* not synonymous. Spelling variants should be grouped together. This sometimes requires judgment, e.g., *Automated* and *Automation* are not spelling variants. (For a more detailed discussion see Section F2.4.1.) It is advantageous to disambiguate homo-

nyms in this step so that, for example, *Banks (economics)* and *Banks (waterways)* do not get merged by mistake on one card (Compare the discussion on homonyms in Section F2.4.2).

With the exceptions mentioned at the end this step can be performed by clerical staff or computer. If Step F2.2 is performed as a manual procedure employing judgment, the critical problems can be resolved there.

F2.2 First Round of Merging: Merge Information for Identical Terms

In the previous step the cards have been grouped into packages. Within each package we have identical or nearly identical terms or terms starting with the same word. In a second step the information contained on the cards of each package is merged on one card (record), as illustrated in Figures 56–58.

F2.2.1 Procedure for merging cards and keeping track of sources (technical)

With manual procedures there are two possible places to put the record that results from merging: one may put it on a fresh card, or one may select a card already in the package and transfer only the information from other cards, thus saving work.

The following criteria may be used in selecting a card (listed in decreasing priority):

(a) Select the card that contains the largest amount of text (e.g., a definition). This will minimize the work needed for the transfer of information from other cards.

(b) Select the card that has been made up from a preferred source. A preferred source may be a thesaurus using structural principles similar to those to be used in the thesaurus to be developed.

(c) Select the card that is most legible.

If a nonselected card contains a lengthy definition, one may just clip it onto the selected card and establish the proper link by a circled number.

While merging cards, one has to keep track of the sources as follows (we repeat here the process already described in Section F0.7.2(2)): Assume that card 1 is the selected card and that card 2 is the card from which information is to be entered on the selected card in the operation of merging. In the crude form all one has to do is to enter the source indication from data field MT of card 2 into data field MT of card 1. If one wants to keep track of spelling variants, then one has to check first to see whether the Main Term on card 2 is a spelling variant of the Main Term on card 1. If so, the Main Term from card 2, together with its source indication, is entered into the field SP of card 1. For ST, BT, NT, and RT, one simply checks for each term given on card 2 to see whether it is already given on card 1. If so, noth-

ing needs to be done. If not, the term is added to card 1. If the detailed form is to be used, things are more complicated. Before any information is transferred from card 2 to card 1 the source code given in data field MT is added to every term entered in any other data field of card 1. When the information from card 2 has been transferred one proceeds as follows. First the source indication (with underline) from data field MT of card 2 is entered in field MT of card 1, (or the Main Term from card 2, together with the source indication, is entered into field SP of card 1) as before. The new feature in the detailed form is that sources are given for data elements in other data fields too. Let us explain this using as an example a term in data field RT of card 2. If the term is already in the data field RT of card 1 only the source code from card 2 is added to the term. If the term is not yet contained in the data field RT of card 1, then it is entered there together with the source code from card 2. (Source codes in data fields other than MT, SP, and ST are not underlined.) An example is given in Figure 56. Further examples are given in Figure 57. Note that in example 3 the term *Attorney, lawyer* is treated as a synonym of the term *Lawyer* (this is done only during the construction phase). An example of merging on a thesaurus form is given in Figure 58.

F2.2.2 Steps after the first round of merging

In most cases one may proceed after this to Step F3, "Work out the structure of the thesaurus". However, there are two exceptions:

(1) Sometimes there is a big thesaurus or other prearranged source that cannot be included in the term collection at the beginning but that could supply useful information for the terms that have been collected from other sources. In this case one should consider "pulling" this information, as described in F2.2.3. This is particularly useful in small projects where an exhaustive collection of terms and relationships is not possible.

(2) In the first round of merging, nothing is done about synonyms. If the area of the thesaurus is not too complex and interrelated and if it may be divided into subject fields and subfields without too much overlap, synonyms will be detected later on in Step F3, "Work out the preliminary structure of the thesaurus". Synonymous and Equivalent (quasi-synonymous) Terms are very likely to be sorted into the same subject field and subfield in this case. If, on the other hand, the area of the thesaurus is complex and interrelated and not easily subdivided, Synonymous and Equivalent Terms are likely to be scattered over different subject fields and subfields during the sorting in Steps F3.1 and F3.2, and there is the danger that the synonymity will never be detected. Therefore, a second round of merging, to be described in Section F2.3, is recommended in this case. In this second round of merging, the information contained (after the first round of merging) in data field ST and possibly in data field SP (spelling variants) is used to bring together the records for Synonymous and Equivalent Terms. The procedure is rather intricate and cumbersome. It is not recommended unless it is really necessary.

INFORMATION RETRIEVAL

WRO800	LB05	AV	AZ	BY	AR15	DD0502
BR	EJ	CM	HI	EI	LM1506	FC
MS	IE	MR11	MZ05	NAO010	NE	NO
SP	VO					

DD	SNO	The use of computers, electronic accounting	
	1	machines, and similar mechanical devices to	
	2	organize store and retrieve recorded	
	3	information. For the use of manual	
	4	techniques in such activities see	
	5	(documentation).	
	9	DD FR 990 Frequency of term in DDC-collection	
	USE	Information storage and retrieval	BY
	UF	Document retrieval	EI
		Information storage and retrieval	MZ
		Library searches	HI
		Records retrieval	EJ
		Records retrieval	HI
	BT	Data processing	AV
		Documentation	DD
		Documentation	WR
	NT	Computerized information retrieval	AV
		Data bank	AV
		Data processing	SP
		Data recording	SP
		Data retrieval	WH
		Document retrieval	WH
		Environmental information retrieval	AV
		Information dissemination	AV
		Information storage	AV
		Search structuring	WH
		Stinfo	AV
		Vocabulary development	AV
	RT	Bibliographies	WR
		Computers	BS
		Data collections	WR
		Data retrieval	EJ
		Documentation	EJ
		Electronic accounting machines	BS
		Filing systems	EJ
		Image storage	SP
		Index terms	EJ
		Indexes (locators)	EJ
		Library sciences	EJ
		Machine translation	BS
		Microfilm	EJ
		Microfilm selectors	EJ
		Publications	WR
		Records management	EJ
		Records storage	EJ
		Search questions	EI
		Selective dissemination	EI
		Translations	WR
		Indexing vocabulary	WH

This example is from the development of TEST. The two-letter codes stand for sources: 23 sources contained the term, and they contained further information as shown in the different fields. DD0502 means that in the DDC thesaurus the item is assigned to COSATI field 05, group 02. Note that in UF two lines could be replaced by

RECORDS RETRIEVAL	EJ,HI

and in BT

DOCUMENTATION	DD,WR

Figure 56. Merging of data elements from different cards for the same term (F2.2). (Source: Heald 1967, issued by the Office of Naval Research, Department of Defense.)

	Card No.	Entry on card	Comments
		Example 1	
Before merging	1	B22.cl Army (1m)	Notation after first source code omitted; assumed to be the notation that precedes the term
	2	BGH Army (2b)	
	3	15.20.1 Army (3)	
After merging	3′	Merged on card 3: 15.20.1 Army (3) (1m: B22.cl; 2b: BGH)	':' separates source code from notational symbol in that source. ';' separates different sources
	4	or merged on new card Army (1m: B22.cl; 2b: BGH; 3: 15.20.1)	
		Example 2	
Before merging	1	474 Attorney (1)	
	2	338 Attorney (1b2)	
	3	K51 Attorney, lawyer (2c)	
After merging	2′	338 Attorney (1b2) (1:474) ST Attorney, Lawyer (2c:K51)	Card 2 contains a definition, therefore merged on card 2 Different form of term in source (2c) treated as synonym
		Example 3	
Before merging	1	5.A.d Parliament (1h)	
	2	5.B.d Parliament (1h)	
	3	453 Parliament (2a)	
	4	453 Parliament, control of executive branch (1c)	
	5	G51 Parliament, legislative assembly (2c)	
	6	I42 Parliament, legislative assembly (FR)	
	7	19.83 Parliament, Parliamentarianism (3)	
	8	452 Parliament, senate, committees (1)	
After merging	3′	Merged on card 3 453 Parliament (2a) (1h:5.A.d;1h:5.B.d) ST Parliament, control of executive branch (1c:453) Parliament, legislative assembly (2c:G51; FR:I42) Parliament, Parliamentarianism (3:19.83) Parliament, senate, committees (1:452)	

Figure 57. Further examples to illustrate merging in the first round (F2.2).

Figure 58. Example of result of merging on a thesaurus form in the first round (F2.2). (The detailed form of keeping track of sources has been used. In the crude form only the source codes that are underlined would remain. The entry in source AR has been added to illustrate the merging procedure; it would not normally be detected in the first round.)

Entry in Source TH

659.5 Color TV picture tube
ST Color kinescope
BT 435.7 Color TV receiver
NT 478.2 Color TV screen
RT 568.3 Flat picture tube
075 Radiation hazard

Entry in Source AR

Color kinescopes
UF Color TV picture tubes
BT Kinescopes
Color TV receiver
NT Color TV screen

Entry in Source SK

Color TV picture tubes
UF Color television picture tubes
BT Color TV receiver
NT Lawrence tubes
Shadow mark tubes
RT Radiation hazards

Entry in Source KL

TC904 Color television picture tubes
BT TK25 Color television set

Result of Merging

0	1	2	3	4	5	6		check hierarchical level	02 check type:	DS	OP	NP	EL	CH	03 Subject Field

05 Notation: _____ 10 MT: Color TV picture tube (TH:659.5) (AR;SK)

12 Stand. abbr. (AB): _____

20 Spellings (incl. abbr.): Color television picture tubes (KL:TC904;SK)

30 Synonymous T. (ST) (incl. equiv. t.): Color kinescope (TH) (AR); Color TV display tube (TH)

4 Classification:

42 Category (CA): _____

44 Semantic factors/Broader Terms (BT): Color TV receiver (TH) (AR;SK); Kinescopes (AR); Color television set (KL)

45 Is semantic factor of Narrower T. (NT) Color TV screen (TH) (AR); Chromatrons (AR); Shadow mask tubes (AR;SK); Lawrence tubes (SK)

46 Related Terms (RT): Flat picture tube (TH) (SK); Radiation hazard (TH)

50 Translations (TR):

F: _____
G: _____
R: _____
S: _____

60 Definition, scope note (SN):

70 Unspec. rel. (UN): _____

81 Editor/Date: _____

If neither (1) "Pulling" nor (2) "Second round of merging" apply the reader may turn immediately to Section F2.4.

F2.2.3 "Pulling" information from additional sources (match and merge)

In addition to the information for a term merged from the cards prepared in Step F1 one may look up the term in a big thesaurus (for example, TEST), a big dictionary, or other sources and add the information given there (match and merge). The term itself as well as the entry for the concept involved has to be found. If the Main Term given in the file cannot be found in the big thesaurus, one should try Synonymous and Equivalent Terms given in field ST or spelling variants given in field SP. If the term finally found in the big thesaurus is not a preferred term, follow the SEE ST or USE ST instruction given to obtain the entry for the concept involved.

Example:

After 1st round of merging we have

Cyclophones

UF ST Additrons.

Looking for Cyclophones in the big thesaurus, we find nothing. Therefore, we look up Additrons. There we find

Additrons

USE ST Trochotrons.

Therefore, we look up Trochotrons and find the entry for the concept involved. This record gives, for example, the Broader Term Counting tubes.

The information taken from the big thesaurus may be grouped into three types:

(1) *Terminological information.* New synonyms for a term may be given (these synonyms may be terms already contained elsewhere in the file or terms new to the file). One can also note which term has been selected as the preferred term in the big thesaurus and copy that decision.

(2) *New BT, NT, and RT relationships* between concepts already represented in the file. A special case in point is the introduction of finer distinctions in these relationships. For example, a source may put together into one field "see also" both NT and RT. The information from the big thesaurus can be used to distinguish between NT and RT.

(3) *Entirely new concepts.* In particular one should take care to include in the file all concepts that are broader than any concept in the file.

,1 Procedure for "pulling" (technical). To obtain new BT, NT, and RT relationships and entirely new concepts one proceeds as follows (manual procedure employing judgment; for computer procedures see Section G2.2.1).

The card for the Main Term *A* is being compared with the entry found

in the big thesaurus. Check each Broader Term (Narrower Term, Related Term) given in the big thesaurus to see whether it is already on the card for *A*. If the BT, NT, or RT cross-reference is not given on the card for *A*, add it in the appropriate data field and check whether or not a card for the added term or a synonym is already contained in the working file. If the added term is not contained in the working file already, one should consider including it. If the new term is a Broader Term, it should always be included. If the new term is a Narrower Term or Related Term, a decision has to be made as to whether the new term will be useful in the thesaurus to be constructed. A new term is included in the working file by transferring the entire entry from the big thesaurus onto a card. In the case of a Broader Term one should check whether it, in turn, has Broader Terms that are not yet contained in the working file. If so, the entries for these Broader Terms have to be pulled as well, and so on. The same procedure could, of course, be followed for Narrower and Related Terms, but this would lead too far. Whenever a whole entry has been pulled from the big thesaurus, the "starting term" is marked so that one knows later on why the term has been pulled. If the big thesaurus contains a cross-reference to a term not to be included in the working file, the cross-reference is not included in the working file either.

The whole process may be performed either as merging in the first round is performed for each term or as a separate step after the merging has been done for all terms in the working file.

An additional note is necessary. As long as the Main Term in the working file is the same as the Main Term in the entry being pulled from the big thesaurus, pulling corresponds to merging in the first round. But whenever we look for a synonym that occurs in the working file card and/or follow a USE instruction in the big thesaurus, we are making use of the USE instructions that are taken from the sources and included in the working file and/or the USE instructions given in the big thesaurus. This corresponds to merging in the second round, to be discussed in the next section. In doing so, we are dependent on the quality of these USE instructions. The problem of prior editing occurs in pulling as well as in the second round of merging (see Section F2.3.3).

In order not to complicate this description too much, keeping track of the sources has not been considered so far. It is rather simple: In the crude form, enter the source indication for the big thesaurus after the term that is the Main Term in the big thesaurus (this term can be MT, SP, or ST in the working file). In the detailed form underline the source code after this term. Furthermore, enter the source code for the big thesaurus after the appropriate terms in all other data fields too. (For a more detailed description of the procedure, see Section F0.7.2.)

F2.3 Second Round of Merging: Merge Information for Terms in the Same Concept Class (Advanced and Technical)

The second round of merging is necessary if and only if the area of the thesaurus is complex, interrelated, and not easily subdivided. The second round of merging makes use of the information contained in the field ST (and possibly SP) after the first round of merging.

F2.3.1 The procedure (algorithm)

Basically what we want is this: Given a file like the one depicted in Figure 59 and all the synonyms for each term in the file. Create one entry (card) in which all the information given for each of the synonyms is merged. Delete all entries that are then obsolete. This is achieved in two passes through the file by the following algorithm which may be performed either manually or by computer. An example to illustrate the algorithm is given in Figure 59, a flow chart in Figure 60, and an example with actual terms in Figure 61.

Pass 1: Start with *A*. Look up *D*, flag *D* "to be deleted", add "ST* *A*" and merge information from *D* to *A*. Look up *F* and do the same. This brings

	Original	Added through algorithm
A	ST D,F,K	ST J,L,N,P,U
B	ST H	
C		
D—		ST * A
E		
F—	ST A,J	ST * A
G		
H—		ST * B
I		
J—		ST * A K — ST * A
L—	ST A,N	ST * A
M		
N—		ST * A
O		ST S
P—	ST D	ST * A
Q		
R		
S—	ST O	ST *O
T		
U—	ST K	

Figure 59. Sample file for the second round of merging (F2.3.1). (Note Minus-sign "—" is flag for entry to be deleted.)

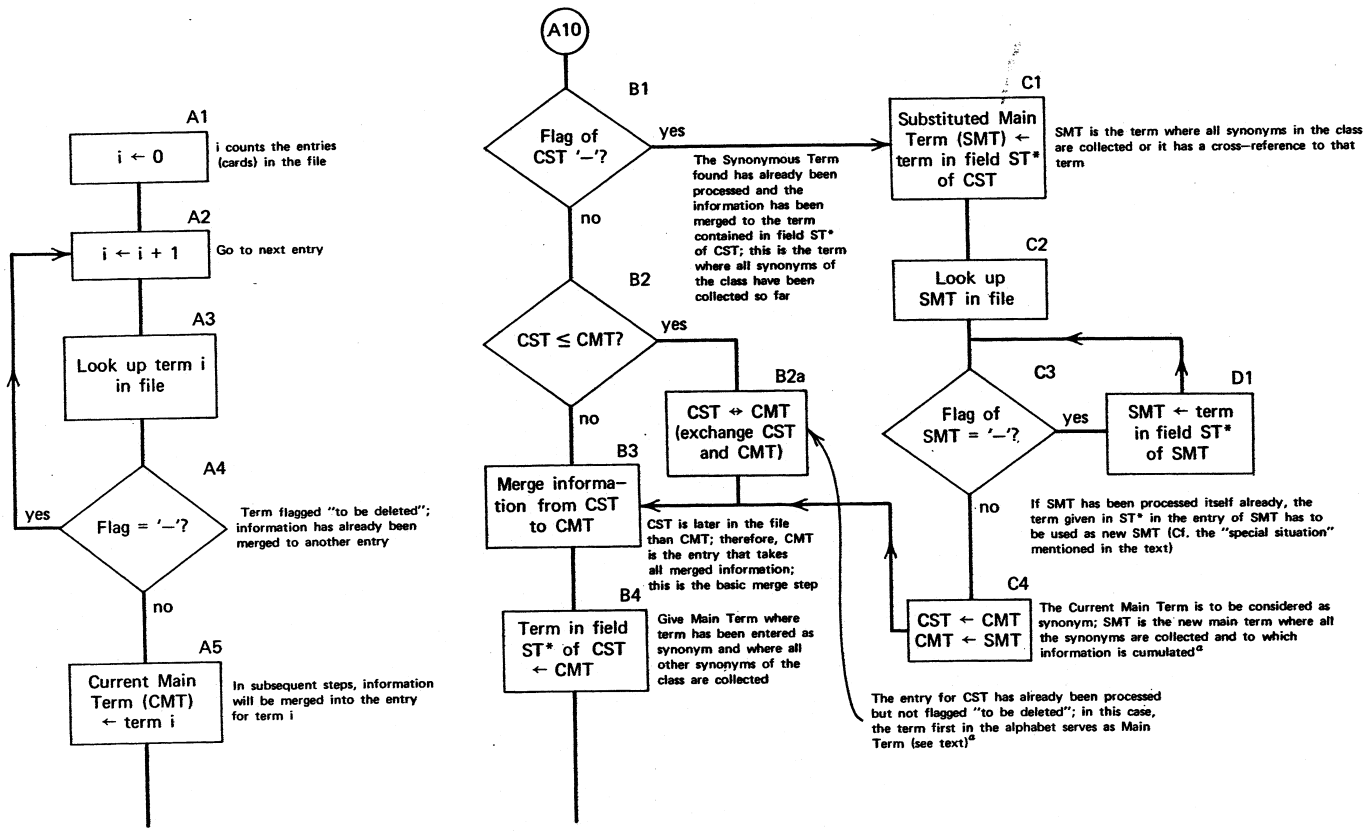
a new term into data field ST of *A*, namely *J*, added at the end. The next term to be looked up is *K*. It is not found in the file, and a new entry *K* ST* *A* is created and flagged "to be deleted" (this will cause *U* to be picked up as synonym of *A*; in a manual procedure one might omit this additional cross-reference at the expense of not picking up *U* as a synonym.) Next look up *J*, . . . Now field ST is exhausted; therefore, proceed in the list and process *B* in the same way. Coming to *D*, the flag "to be deleted" is detected and *D* is therefore skipped. The same is true for *F* and *J*. *L* does not have a flag when encountered first. However, when looking up *A* it is detected that *A* has already been processed and is not flagged "to be deleted". Therefore, *L* is flagged "to be deleted", "ST* *A*" is added and all the information is merged to *A*. This situation occurs if a term has an ST cross-reference to another term preceding it in the alphabetical sequence. *L* is added to the field ST of *A* and marked as processed. *N* is also added to the field ST of *A* in the merging procedure. *N* then has to be processed in the same way as *D*, *F*, and *J*. (Note that one could just as well add all information to the entry for *L* and flag *A*. Choosing routinely the term first in the alphabet is convenient in a computer program, especially with respect to keeping track of which synonyms have already been looked up. But it is less desirable in a manual procedure where the information should be added to the entry that already has more information.) If field ST of *A* is exhausted again, go on to *M* and continue. *P*, like *L*, does not have a flag when encountered first. However, when looking up its synonym *D*, the flag "to be deleted" is detected and "ST* *A*" is found. The same action as in the case of *L* is taken. (In this case, the synonymity between *A* and *P* is detected due to the fact that they had the Synonymous Term *D* in common.) This situation shows that it would not be appropriate to delete *D* before pass 1 is completed. When coming to *U* look up *K*, find *K* ST* *A*. Therefore, transfer all information from *U* to *A* and enter ST *U* with *A* and ST* *A* with *U*.

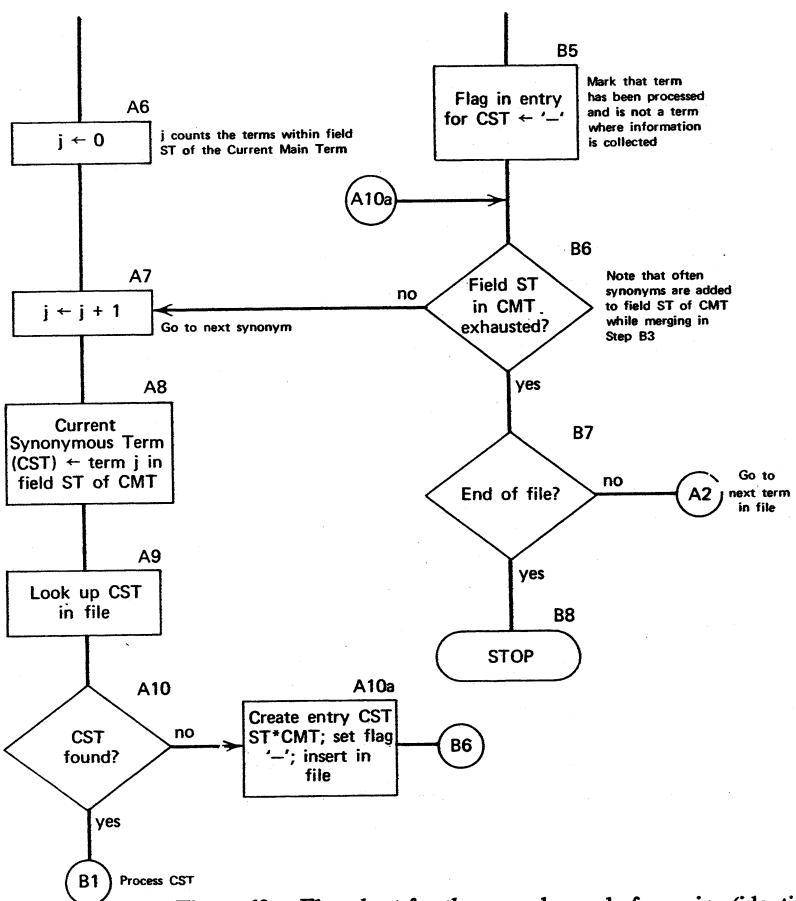
A special situation, not shown in the example, may also arise.

Example:

A	*ST D,F*
K	*ST U*
P	*ST A,K*
U	

While processing *K* the information from *U* is transferred to *K*, and with *U* the cross-reference ST* *K* is entered. In processing *P*, *K* is transferred to *A* into data field ST, and the next step is to look up *K*. In this case the flag with *K* is changed to "to be deleted", ST* *A* is added, and the information is cumulated to *A*. A double transfer of information, from *U* to *K* and then from *K* to *A*, is necessary. (In the case of *L* double transfer of information—





Legend: CST ← CMT, the value of the variable CST is set to the current value of CMT, in other words, the term in CMT is now also in CST;

CMT, Current Main Term, the term where all synonyms of a class are collected and where the information from their entries is merged;

CST, Current Synonymous Term, that synonym in field ST of the Current Main Term that is now being processed;

SMT, Substitute Main Term, becomes Current Main Term unless flagged "to be deleted"; if SMT is flagged, the term given in field ST* of the (old) SMT becomes the new SMT.

(a) Note: All the synonyms given in field ST of term i, the original Main Term, are transferred to the new Main Term in Step B3. The further processing (Step B4 and following) is done for the new Main Term, and all synonyms transferred from term i are processed then. The count for i is not changed, however, so that term i + 1 is processed after B7 has been reached.

Figure 60. Flowchart for the second round of merging (identifying classes of synonyms) (F2.3.1).

from *N* to *L* and then from *L* to *A*—is avoided by processing the terms in alphabetical order.) Also, if we have *Z* ST *U*, we are referred from *U* to *K* and then from *K* to *A*

Pass 2: Delete all records flagged "to be deleted".

In manual processing the flag "to be deleted" may consist of a paper clip put on the card and the indication "ST* *A*" may be achieved by entering A in the field ST and underlining it. Cards with main records do not have a paper clip.

The algorithm described is a natural way to identify groups of synonyms in a manual procedure. Since the general problem of identifying equivalence classes in a set starting from binary equivalence relationships (of which our problem is a special case) occurs fairly often, it is quite possible that, for processing by computer, better algorithms can be found in the computer science literature.

In order not to complicate the description, details of the actual process of merging two entries and keeping track of the sources have been omitted. Refer to Sections F0.7.2 and F2.2.

F2.3.2 Treatment of terms that consist of a string of Synonymous Terms

Sometimes a term (either the Main Term in data field MT, or a synonym in data field ST) consists of a string of synonymous or quasi-synonymous words or phrases, separated by commas; for example: *Beer, ale, malt liquor.* In this case the constituents are considered to be synonymous or quasi-synonymous to the term as a whole and to each other. Therefore, starting from *Beer, ale, malt liquor,* one should look up *Beer, Ale,* and *Malt liquor* (unless one of these terms has been processed already). In a manual procedure in which a certain capability of judgment can be assumed (see below) appropriate instructions should be given that the constituents are to be treated in the same way as synonyms. (However, there are cases where the terms in a string are not synonymous; see Section F2.3.3,2.) In a purely mechanical procedure (manual or computer) the constituents should be entered into field ST prior to performing the second round of merging. The procedure is further illustrated by the examples given in Figure 61.

Of course, this procedure is not applied to multiword terms like *Gross national product* or *Electron tube* (note the absence of commas!).

F2.3.3 Editing during or prior to the second round of merging

In the algorithm described above the appearance of one wrong synonym in a source may lead to the merger of two whole series of records that actually belong to different concepts and should be kept separate. It is rather awkward to disentangle such a mess afterwards. Therefore, an editor has to exercise judgment as to what terms appearing in field ST (and possibly in field UN Unspecified relationship) should be used in the second round of

merging, and what terms should not. If the second round of merging is performed manually by a person capable of making such judgments, the editing can be done during the second round of merging. If the second round of merging is performed in a merely mechanical way, especially if it is done by a computer, prior editing is necessary. The following points have to be considered in editing:

(1) Wrong synonyms are especially likely to come from sources that do not distinguish between UF ST and UF NT (use the crude lead-in form).

(2) If artificial synonyms are created from string terms as described in F2.3.2, careful editing is necessary. E.g., in *Parliament, parliamentarianism, Parliamentarianism* is not a synonym for *Parliament.* Also sometimes two specific terms are strung together to designate a broader concept, such as *Roughness, smoothness.* Again the procedure is not applicable.

(3) Often a term that is wrong as a synonym is useful if specified in some other kind of relationship. A simple change in the relationship indicator will do in this case. The term may come from a source, or it may be an artificial synonym. Whereas *Parliamentarianism* is not synonymous to *Parliament,* it is a useful Related Term.

(4) The problem of spelling and morphological variants described in Section F2.4.1 can be taken care of in this step as well by entering, for example, the relationship *Filtering* Spelling Variant *Filtration.*

(5) Homonyms can also be detected in this step of editing, as described in Section F2.4.2.

(6) If it is necessary to keep track of the sources precisely, two source codes are used for the relationships newly entered in editing: the code of the source that contributed the information being edited and the code for the editor.

(7) In a cumulative thesaurus special problems arise, as described in Section K1.3.

F2.3.4 Concluding remark

In the second round of merging, groups of synonyms and quasi-synonyms are detected in a purely mechanical way, based on ST cross-references in the sources. To keep the procedure simple the cumulated main record is kept under that term of a group that comes first in the alphabet. This in no way prejudices the selection of the preferred term. There is no guarantee that any group of synonyms is complete, especially if the ST cross-references in the sources are not very well developed. Further synonyms will then be detected later in Step F3, "Working out the preliminary structure of the thesaurus".

F2.4 Remarks Regarding Both Rounds of Merging

F2.4.1 Spelling and morphological variants

In both procedures spelling variants should be treated as identical. This does not present a problem with manual procedures as long as the spelling vari-

Figure 61. Examples illustrating the second round of merging (examples 2 and 3 from Figure 57 are continued into the second round of merging) (F2.3.1).

	Card No.	Entry on card	Comments
		Example 2	
After 1st round of merging before 2nd round of merging	2′	338 Attorney (1b2) (1:474) ST Attorney, Lawyer (2c:K51)	
	4	327 Barrister, attorney (7) (3:15.72)	
	5	U25 Lawyer (FR) ST Solicitor	
After 1st step in second round of merging → : added in first step	2″	338 Attorney (1b2) (1:474) ST Attorney, lawyer (2c:K51) → Lawyer (FR:U25) → Solicitor (FR:-U25)	Step 1: *Lawyer* appears on card 2′. Therefore, look up, find card 5, transfer into 2′ resulting in 2″, tag 5 "to be deleted" (5—), add ST*Attorney. Now *Solicitor* appears on card 2′. Therefore look it up. No entry found, Step 1 finished. Go to next card. -U25 as notation for *Solicitor*, because it is not the preferred term in FR.
	4	327 Barrister, attorney (7) (3:15.72)	
	5—	U25 Lawyer (FR) ST Solicitor → ST*Attorney	
After 2nd step in second round of merging → : added in second step	2‴	338 Attorney (1b2) (1:474) ST Attorney, Lawyer (2cK51) Lawyer (FR:U25) Solicitor (FR:-U25) → Barrister, attorney (3:15.72;7:327)	Step 2: In due course, one arrives at card 4. Attorney is looked up and card 2″ found. Since 2″ has been processed and is not tagged "to be deleted", the information from 4 is transferred on it, card 4 is tagged "to be deleted" and "ST*Attorney" is added. Note that the original cards 1-5 can be completely reconstructed from card 2‴.
	4—	327 Barrister, attorney (7) (3:15.72) → ST*Attorney	
	5—	As above	

		Example 3	
Merging in second round performed until here	9	375 Congress, legislative assembly (4) ST Legislative assembly (2b : IEF)	
	10—	IEF Legislative assembly (2b) ST*Congress, legislative assembly	
Merging in second round still to be performed in these cards	3′	453 Parliament (2a) (1h:5.A.d;1h:5.B.d) ST Parliament, control of executive branch (1c:453) Parliament, legislative assembly (2c:G51;FR I42) Parliament, Parliamentarianism (3:19.83) Parliament, senate, committees (1:452)	The process continues with card 3′. All terms recorded there are looked up. This leads first to card 10—. This card is flagged, and leads, in turn, to 9. Since 3′ contains more data, it is retained as main card and information from 9 is transferred, 9 is flagged, etc. Next we find card 11, and transfer of information results in the last data element on card 3. Note the link between 9 and 3 established by *Legislative assembly.*
	11	414 Parliamentarianism (1c)	
Merging in second round performed for all cards of example	9—	375 Congress, legislative assembly (4) ST Legislative assembly (2b:1EF) → ST*Parliament	
	10—	IEF Legislative assembly (2b) ST*Congress, legislative assembly	
	3″	453 Parliament (2a) 1h:5.A.d;1h:5.B.d) ST Parliament, control of executive branch (1c:453) Parliament, legislative assembly (2c:G51;FR:I42) Parliament, parliamentarianism (3:19.83) Parliament, senate, committees (1:452) → Congress, legislative assembly (4:375) → Legislative assembly (2b:IEF) → Parliamentarianism (1c:414)	Now, not all results of this mechanical procedure are useful. *Congress* is the legislative assembly for the U. S. and should have a separate entry. Parliamentarianism *not* a synonym of Parliament (source 3 is wrong) and should also retain its main entry. These are decisions to be made by the lexicographer who will also note that *Parliament* is a semantic factor of *Congress* and a Related Term of *Parliamentarianism.*
	11—	414 Parliamentarianism → ST*Parliament	

ants are near neighbors in the alphabetical sequence. It does present a problem, however, in computerized processing.

There is a more fundamental problem, illustrated by pairs such as *Filtering* and *Filtration, Safe* and *Safety, Automated* and *Automation.* From a morphological point of view these are clearly spelling variants. From a semantic point of view the terms in each of these pairs might be sufficiently different to justify treating them as different descriptors or at least keeping one as a Synonymous or Equivalent Term. Intellectual judgment is necessary to make these decisions. (Compare Section C6 on the functional distinction between Synonymous Terms and Spelling Variants.)

F2.4.2 Homonyms

In all the procedures described (first round of merging, pulling from a big thesaurus, second round of merging) there is the danger that cards for homonys might be merged. In a manual procedure employing judgment this danger can be avoided easily enough once one is aware of it. In mechanized procedures things are more difficult. There is no way to avoid the merger of records for a homonymous term unless it is explicit from at least one source that the term is homonymous. In this case all records for this term can be printed out for decision by an editor. Otherwise, the homonymy must be detected later on and the wrongly merged records must be disentangled. If the first round of merging has been done in a purely mechanical way, editing should take place before the second round of merging, as discussed in Section F2.3.3. In this step it should be easy enough to detect records created by wrongly merging the records for two different meanings of a homonym. Obviously, the result of such a merge is odd, especially as there are likely to be many wrong synonyms.

F3 WORK OUT THE PRELIMINARY STRUCTURE OF THE THESAURUS: THE SYNONYM-HOMONYM STRUCTURE, THE EQUIVALENCE STRUCTURE, AND THE CLASSIFICATORY STRUCTURE. SELECT PREFERRED TERMS

Conceptually, it is very important to keep the distinction between the synonym-homonym structure, the equivalence structure, and the classificatory structure as expounded in Chapters B and C. However, in the practical development of a thesaurus problems of all three levels have to be considered in one and the same procedure. Synonymous and Equivalent (quasi-synonymous) Terms are scattered over the whole alphabet. Therefore, terms must be sorted according to a preliminary coarse classification so that groups of Synonymous and Equivalent Terms can be detected. This is the only way to deal with the terminological problems and to form new concepts by consoli-

dating equivalent concepts. The procedure essentially consists of a "cascade-type" sorting of terms, first into broad subject fields, then into subfields, and then into small groups of terms each corresponding to an ISAR concept.

The procedure to be described can be viewed best as the interaction of two principles: (1) the "deductive" principle: start from broader concepts and subdivide them further and further according to some preselected viewpoints, arriving at specific concepts; (2) the "inductive" principle, start from specific concepts, arrange them in small groups which correspond to less specific concepts, arrange those in groups, and so forth, finally arriving at the very broadest concepts. In most practical situations, the inductive principle plays a larger role, but the interaction of both principles is always necessary.

The procedure described in the following leads to a working file in classified order. In the Roget-Soergel model this corresponds to the arrangement of descriptors in the main part. In the TEST model the main part is arranged alphabetically, and it might seem, therefore, that the procedure is not appropriate for the development of a TEST-like thesaurus. However, this is not so. For the development of the thesaurus structure, the working file should be arranged in classified sequence in any case. It will become clear in the following sections that classified arrangement is essential for a reliable detection of synonyms, for the proper definition of concepts, and for uncovering their interrelationships. Also, classified arrangement makes it much easier to discuss all terms belonging to a certain subject field with an expert in that subject field. The user version of the main part, in which terms are arranged alphabetically, should then be produced in Step F5.7 (see Section F5.7.3).

F3.1 Define Broad Subject Fields and Sort Terms into These Broad Fields

By looking at the material collected in the previous step one should get some idea of what the subject fields should be. Further information can be gathered by looking at the major divisions of existing classification schemes, or at tables of contents of textbooks or similar documents. Further clarification, especially concerning the delineation of the different subject fields, can be achieved by asking subject experts, e.g., by organizing a discussion, as described in Section F0.3.3. Terms are sorted into these broad subject fields. In the course of the development of the thesaurus a partial or complete reshuffling of the subject fields may prove necessary. If a thesaurus is developed by parallel development of constituent thesauri, the subject fields are given by the general framework of the total thesaurus (the "umbrella classification"), see Section K2.3.1. Instead of subject fields one may also choose facets as the primary subdivisions. In Step F3.2 subfacets must then be defined.

F3.2 Define Subfields within Each Subject Field and Sort Terms Accordingly

The broad subject fields are now subdivided into smaller but still sizeable subfields, and the terms are sorted into these subfields. The remarks on how to obtain suitable subject fields apply to subfields as well. However, it is usually sufficient to consult subject experts according to the procedure described in Section F0.3.2; discussions would involve too much effort for the purpose at hand. Quite often a major reshuffling of these subdivisions will prove necessary later in Step F4, "Work out first draft of the classified index".

Notes on F3.1 and F3.2:

Sorting can be done in both steps as a two-step process: A professional writes the code for the subject field or subfield on the card for the term or encircles the appropriate code as preprinted on the card. The actual sorting is then done by clerical staff (or by a computer).

If an appropriate list of fields and subfields can be drawn up from the beginning, Steps F3.1 and F3.2 may be performed at the same time. One may use the checklist technique in analyzing terms. This may be assisted by providing a form such as the one given in Figure 62. In essence this form gives the outline of a faceted classification, and terms are analyzed according to a faceted scheme. We could also say they are decomposed into semantic factors on a very broad generic level. Note that some of the downward arrows in Figure 62 correspond to autonomous subdivisions (such as the arrow going down from *Materials*). Other arrows correspond to subdivisions according to another facet; for example, the arrow going down from *Supplies*.

Step F3.1 or Step F3.2 is also the appropriate time to fill in the information in the data fields 41 Subject field and 42 Facet (of Figure 21, Section C7) if those data fields are to be used in the thesaurus to be built.

F3.3 Work Out Detailed Thesaurus Structure. Select Preferred Terms. Merge Information for Terms in the Same Concept Class

Each of the subdivisions created in the previous step contains only a limited number of terms. These terms can be kept in mind or displayed all at the same time so that it is possible to detect relationships among them and to work out the detailed structure of these relationships.

The elaboration of the detailed structure can be performed in two steps, as described in the following Sections F3.3.1 and F3.3.2. However, the functions to be performed in both steps cannot be completely separated, so that sometimes one has to go back and forth between the two steps. Sections F3.3.3 and F3.3.7 contain additional considerations to be taken into account in one or both of these steps.

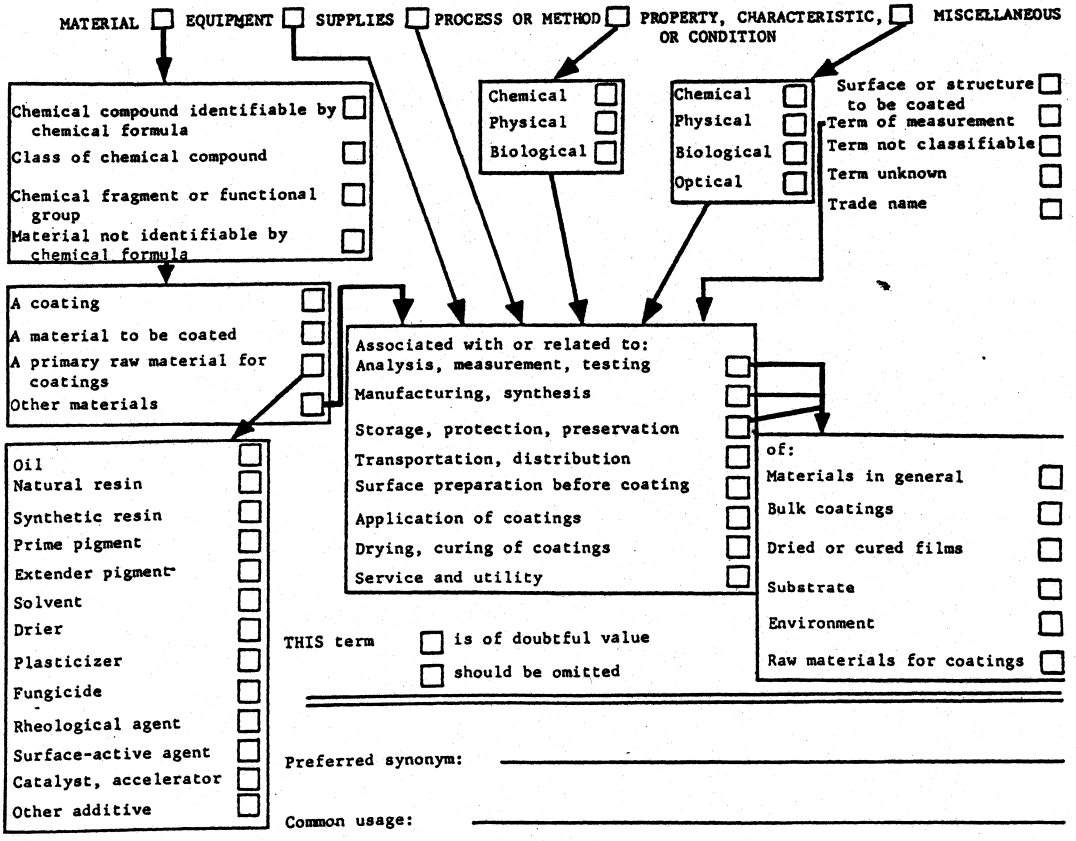


Figure 62. "Road map" for the analysis of terms (Section F3.2). (Courtesy American Society for Information Science.)

During the elaboration of the detailed structure it is useful to consult subject experts on specific topics according to the procedure described in Section F0.3.2.

F3.3.1 Work out the synonym-homonym structure and the equivalence structure

The cards of Synonymous and Equivalent Terms are grouped together. In principle this is a continuation of the "cascade-type" sorting procedure started in Step F3.1 and continued in Step F3.2. There is one difference, however: In these former steps the subject fields and subdivisions were established before the sort began. Here the groups are established during the very sorting procedure, as follows: The first card is laid on the table, thus "opening" a group. If the second card contains a term synonymous or equivalent to the term on the first card, the second card is added to that group; otherwise, a new group is opened, and so forth. All cards of a subdivision are processed in this way, that is, either added to a group already available or used for opening a new group. The viewpoints to be considered in the formation of groups of Synonymous and Equivalent Terms have been dealt with in Sections B1 and B2. Synonyms given on the card may also be helpful in the process (unless the second round of merging has been performed, in which case synonyms on the card have been used already). It is possible and even occurs frequently that a group consists of one card only. The whole process shows how important hierarchy and classified arrangement are for the detection of Synonymous and Equivalent Terms.

For each group a preferred term is selected according to the criteria set forth in Section F0.4.2. It is especially useful to consult experts in order to make correct decisions. The preferred term selected may be any term occurring in data field MT or ST (or SP) on one of the cards. The selection decision is not bound by a selection made in one of the sources (or, in case the second round of merging has been performed, by the completely arbitrary selection there). Often it will be necessary to coin a new term.

The information from all other cards of the group is merged onto the new card for the preferred term. The nonpreferred terms themselves, together with their source indication, are entered in field ST; other information is entered in the appropriate data field (see Section F0.7.2 and F2.2 for the technique of merging and keeping track of the sources). After the information has been merged on the card for the preferred term, all other cards can be eliminated, since they do not contain any information in addition to the card for the preferred term. By this procedure the number of cards is reduced considerably; they can now be surveyed much more easily in the following step.

F3.3.2 Work out the classificatory structure

The previous step was concerned with the elaboration of the synonym-homonym structure and the equivalence structure. Each of the preferred terms now corresponds unequivocally to a concept. We can now turn to the classificatory structure in the set of concepts represented by the preferred terms. In elaborating the classificatory structure, the following questions have to be answered (a full treatment of classificatory structure is to be found in Section C1):

(1) *Is it possible to decompose a concept into semantic factors?* If semantic factors are given in the input, one can draw upon the semantic factors given for different terms in a class; OR parentheses as semantic factors may often be useful to take into account several contributions. Otherwise, or in addition, it is useful to look at the Broader Terms (if any) given for the concept. For example, *Monetary policy* may have the Broader Term *Economic policy* in one of the classification schemes used as a source.

A concept to be used as semantic factor may be already available among the concepts collected so far, but it is also quite possible that semantic factoring gives rise to the introduction of a concept that did not occur explicitly in any of the sources.

Example:

Airports = Air traffic: Traffic stations

Traffic stations is introduced as a new concept (with the editor's code as source code). Of course, a card is made up for the new concept and is filed in the appropriate subject field and subdivision (which need not be the same as the subject field and subdivision just processed). It may turn out later in the process that the concept assumed to be new was available in the collection. This doesn't do any harm, however.

(2) *What hierarchical relationships exist among the concepts?* The best way to indicate many of the hierarchical relationships is to arrange the cards in a linear sequence representing a monohierarchical structure (as described in Section D3.1.1). The hierarchical level (the number of indentions) is indicated on top of the thesaurus form by '+' (for potential descriptors) or '—' (for other preferred terms). Additional Broader Terms of the next higher, and Narrower Terms of the next lower hierarchical level that come to mind are entered on the cards. In many places it will prove necessary to introduce new broader concepts, as discussed in Section C1.4.1 (the reader is urged to reread this section).

(3) *What associative relationships exist among concepts?* In answering these questions, one uses all the information contained on the card for a concept (preferred term).

Notations given—either for the preferred term or for spelling variants

or synonyms—are especially useful in this step: they can be used to look up the concept in the corresponding source scheme; there one can see the whole context in which the source scheme places the concept. Often open-ended sources, such as search requests or abstracts, give very useful hints on the use of a term, on the context in which it occurs, and on relationships to other concepts. Furthermore, an alphabetic index in the KWIC format can be useful (see Section F5.8 for a more detailed discussion). If the terms are in machine-readable form, a KWIC index should be produced for this purpose.

F3.3.3 Use of judgment and creative thinking in processing the information collected from different sources

In steps F3.3.1 and F3.3.2 use is made of all the information on a term or concept as collected from the different sources and merged on one card. However, one is in no way compelled to include all synonyms and cross-references from each source. One may well disagree with a source on a certain relationship. That relationship may then simply be deleted (except in building a cumulative thesaurus, see K1.3). On the other hand, the lexicographer may and should introduce new relationships. If the type of the new relationship is not completely clear, the term is entered in data field 70 UN (Unspecified relationship, of Figure 21, Section C7).

Creative thinking is called for, particularly when new concepts are introduced arising from semantic factoring or needed as broader concepts. This process of concept formation results from the application of a specific way of thinking: The information scientist or classificationist developing a thesaurus is charged with the task of rendering explicit and laying down on paper the structural relationships among the concepts of a field. For achieving this end he applies the tools and the methodology developed in Chapter C. It is therefore not surprising that he sometimes comes up with the formation of concepts that have not been thought of before in that form by the experts in the particular field.

The process of concept formation aims at a complementation of the indexing language in such a way that the subject field in question is completely covered and that overlap between concepts is eliminated as far as this is possible and useful. This activity also leads to the formulation of scope notes and definitions, to be dealt with below. The whole process is continued in step F4, "Work out first draft of the classified index".

This process of concept formation is the essential and truly creative activity in thesaurus development. It is obviously not possible without developing a classificatory structure. (Compare Section C3.1, "Concept formation in thesaurus building").

The following two sections take up two specific problems in this context.

F3.3.4 Introducing more specific concepts

It might occur that one of the source schemes lumps together several related concepts that should be kept separate in the scheme to be developed. Accordingly, new cards have to be made. One may wish to retain the original concept as a broader concept for the newly created ones or one may wish to establish an associative relationship between the newly created concepts.

F3.3.5 Scope notes and definitions

In delineating concepts in step F3.3.1 it is often useful to put down the distinctions explicitly in the form of scope notes or definitions as discussed in Section C3.2. Scope notes are also needed for new concepts arising from semantic factoring or introduced as Broader Terms. In connection with the mutual delineation of concepts, numerous associative relationships will be detected.

F3.3.6 Preliminary selection of descriptors from among the preferred terms

Recall from the summary in Figure 6 (Section B4.2) that a preferred term is the term selected from a class of Synonymous and Equivalent Terms to designate the concept at hand. Only part of the preferred terms are used as descriptors, i.e., in document representations and search request formulations. In the previous steps, we were concerned only with the elaboration of the conceptual and terminological structure. Now we are faced with the problem of which concepts are important enough to be descriptors, i.e., to be included in the indexing language.

In clear-cut cases preferred terms can already be removed from the list of potential descriptors in this step, using the criteria given in F0.4.3 (the final selection takes place after step F5.1):

(1) For some concepts it may become apparent from the structure developed in this step that they fall beyond the scope of the thesaurus and therefore should not be included. The corresponding cards should be taken out of the working file but kept until the thesaurus is finished (one might reconsider some of the decisions).

(2) For less important concepts, especially if they are very specific, one may decide that a broader concept or a combination of concepts should be used in indexing and searching. OP ("other preferred term, nondescriptor") is marked on the corresponding card, and the descriptors to be used are entered in field BT together with a USE instruction. For the purpose of working out the basic hierarchical structure in step F4, one may wish to exclude these cards so that one has to deal only with the smaller set of really significant concepts. The hierarchical level is marked by '-', so that these cards are skipped over in typing the first draft of the classified index in step F4.1. (One could also remove these cards from the

working file and reintroduce them in step F5.1, "Revise main thesaurus file", but this is not recommended.)

F3.3.7 Some suggestions for the technique to be used (technical)

It has been suggested that the following technique gives a better overview during the procedure: for arriving at the groups of Synonymous and Equivalent Terms (step F3.3.1), write the terms belonging to one subdivision on a large sheet of paper in such a way that related terms appear in the same neighborhood, the closeness of the relationship being indicated by the degree of their proximity on the paper. In doing so, one obviously performs at the same time some of the functions of arranging concepts in a meaningful order (Step F3.3.2). Having finished this display, isolate groups of Synonymous and Equivalent Terms and draw a line around them. Within each group select a preferred term and underline it.

A variant of this technique is as follows: the terms are not written immediately on the large sheet of paper but on small slips of paper (2 x 1 cm.) which then can be arranged on a table or pinned onto a board. This technique has the advantage of being more flexible with respect to working out the arrangement of terms.

Methods for the display of relationships between terms by a graphical arrangement have been dealt with in detail in Section D3.

It is questionable whether the application of these elaborate techniques is worthwhile in this phase of thesaurus development. It is quite possible to isolate the groups of Synonymous and Equivalent Terms and to select the preferred terms by the method described in F3.3.1, which uses the index cards that are already prepared. Techniques similar to those described in this paragraph are more appropriate later on, as described in section F4.2.

F4 WORK OUT FIRST DRAFT OF THE CLASSIFIED INDEX (SCHEDULE)

As a result of step F3 one has a very preliminary version of the main part of the thesaurus in a classified arrangement in the form of a file of index cards, the working file. The purpose of the procedure described in this section is to improve and streamline this structure.

F4.0 Classified Index and Cross-References in BT, NT, and RT

A somewhat difficult preliminary point has to be discussed first. A classified arrangement transmits information by the very sequence of terms. As soon as the arrangement is changed, information is lost. Therefore, a preliminary step is recommended: enter the hierarchical relationships for a term as

shown by the classified arrangement on the card for the term so that the classified arrangement can be changed without information loss. The Broader Term on the next higher level and the Narrower Terms on the next lower level should be entered. (One may omit the Narrower Terms since they will be introduced later on anyway as inverse cross-references. On the other hand, it is easier to work with the file if the Narrower Terms are entered from the beginning.)

This remark applies at different points of the procedure to be described. First of all, it should be followed before any changes in the arrangement arrived at in Step F3 are made. The appropriate point is after the preliminary classified index is typed because the Broader and Narrower Terms can be easily seen then. At this point one should also transfer additional hierarchical relationships to be seen from sources that have a classified arrangement (see Section F1.2.2,1. A second obvious point is after the improved classified index has been typed (F4.3). Ideally, one should also keep track of all the changes taking place in rearranging the hierarchy in steps F4.2 and F4.4; however, there are practical limits.

If few changes are expected in the classified arrangement, the effort required for this procedure might not be warranted, and one may follow a procedure otherwise recommended only for thesaurus updating and described in Section J3.2.

Similar considerations hold for Related Terms (Compare Section C1.5).

If detailed keeping track of the sources is necessary, the code of the editor or the code of a source that has a classified arrangement should be used.

F4.1 Type Preliminary Classified Index. Amend Working File

It is not easy to work with the working file in classified arrangement, as produced in F3, because it does not allow for a good overview. Therefore, a classified index should be typed now. The preliminary classified index lists only the descriptor candidates selected in Step F3.3.6, and it gives only the term as such, perhaps supplemented by a notation, but no further information, thus being much shorter and easier to peruse than the working file.

The classified index should be typed as a sequence with indentions. The number of indentions is indicated to the typist by a "+" after the hierarchical level on top of the card. Cards on which the hierarchical level is indicated by "—" and on which OP is marked are skipped in typing. (These are cards for terms that have been ruled out as descriptor candidates for example, because they are too specific. If there is a "—" and DS is indicated, there is an error that should be checked by a staff member.) If one wishes to have extra line spaces, they are indicated by an empty card. Other forms of display are pos-

sible, but usually less suitable at this stage. If such other displays are planned for the user version, they should be drawn up during or after step F4.2 or after step F4.4. (Compare Sections D3.1 and D3.2.)

After the preliminary classified index is available, the Broader and Narrower Terms that can be seen from the classified arrangement have to be entered on the cards, as described in Section F4.0.

F4.2 Improve the Classificatory Structure

The classified index is now copied on cardboard or heavy paper (if the available equipment does not allow for copying on cardboard, the following procedure may be used: type on cardboard, make a normal copy that can be retained, and then proceed as described in the following). Cut the cardboard into small slips containing one term or a group of terms each. Arrange these slips in tree form or as a network (compare sections D3.1.2 and D3.2 as to the format). Since the unqualified tree method needs much space, a modified arrangement is usually to be preferred.

Example:

		Field (Internat. politics)
subfield 1	*subfield 2*	*subfield 3 (diplomatic activities) . . . subfield*
xxxxxxx		*Official visits and other contacts*
xxxx		*State visits*
xxxx		*Visits of VIP's*
xxxx		*Contact of embassy*
xxxx		*with host government*
xxxx		*Exchange of notes*
xxxx		. . .

In this example the preferred terms are arranged within each subfield in a linear sequence with indentions. It would also be possible to carry the tree-type arrangement one level further and then use the linear-sequence-type arrangement. If one chooses this type of arrangement, it is advisable to leave hierarchical subgroups uncut if it seems likely that the elements of the group would be left together in any arrangement. In trying alternative arrangements one can then move the whole group as one block. The group may be cut later of course.

This technique allows one to survey the structure of a whole subject field. Therefore, the classificatory structure, especially the hierarchical relationships, can be checked. By rearranging the slips one can try out different variants of the hierarchical structure and select the best. It is useful to enlist the cooperation of subject experts for the step of trying out the different possibilities for the hierarchical structure.

Further screening as to which preferred terms should be selected as descriptors and which of these should be selected as checklist descriptors may also take place in this step. Decisions should be recorded in the working file, thus preparing for Step F4.4.

During the whole procedure, appropriate BT and NT entries should be made in the working file to preserve information before the classified arrangement is changed, as described in Section F4.0. Additional BT, NT, and RT cross-references may evolve during the process. As a labor-saving device, all these cross-references might be recorded very sloppily in this stage until the notations are available. More thorough recording is then done in Step F4.6. If a new concept is introduced, a card has to be made up. Finally, the working file should be rearranged so as to correspond to the improved version of the classified index.

F4.3 Type Improved Classified Index and Amend Working File

The improved classified index can now be typed and copies can be produced for distribution to subject experts (if numerous copies are necessary, use of a stencil is advisable). In typing, leave enough space at the left margin so that a notation can be entered later on.

If the changing BT and NT relationships were not entered in the working file during the elaboration of the classified index, the BT and NT relationships to be seen from the improved version of the classified index should be entered now.

F4.4 Discuss Classified Index with Subject Experts. Select Descriptors and Checklist Descriptors

The classified index displaying the hierarchical structure is the backbone of a thesaurus. A thorough discussion with subject experts, as described in Section F0.3.3, is therefore in order. Separate discussions should be arranged with experts from each subject field or subfield. In each such discussion the subject field or subfield should be discussed thoroughly, concept by concept, and selected problems from other subject fields or subfields should be dealt with also. In an interdisciplinary approach one might want to discuss the whole scheme with subject experts from different areas as described in section F0.3.4.

It is possible and often useful for the preparation of the discussions to ask for written comments from subject experts as described in Section F0.3.1(2). Copies of the draft of the classified index should be distributed to gather such comments.

The discussion should deal with the following points:

(a) *Does the preferred term represent the concept in question adequately?* This is a terminological problem. If the need arises, recourse can be made to the corresponding card in the working file where all the Synonymous and Equivalent Terms are given.

(b) *Over-all structure of the hierarchy:* selection and delineation of the subject fields and subfields; sorting of the concepts in the subfields; helpful order in the arrangement of concepts on the same level of the hierarchy. However, experience has shown that the subdivision of a subject field into subfields cannot be meaningfully discussed without a more detailed look at the concepts listed within each subfield.

(c) *Individual hierarchical relationships.* In order to make sure that it is correct to indicate *A* as a Broader Term for *B*, ask the following question: while searching for documents on *A*, do you want to retrieve all or most of the documents indexed by *B*?

(d) *Selection of the descriptors* (preferred terms that should be included in the indexing language) and selection of the checklist descriptors (descriptors that are of particular importance in searching and therefore warrant special consideration in indexing). See Section F0.4.3 for the criteria to be applied in the selection of descriptors and checklist descriptors.

(e) *Filling in any gaps* in the indexing language (classification scheme) by introduction of new concepts; new broader concepts are introduced, as discussed in Section C1.4.1, or entirely new concepts are added that up to now have been overlooked.

The resulting modifications are recorded in the draft of the classified index and in the working file. If many modifications have been made, the modified sections of the classified index should be completely retyped. Often it will be useful to repeat the discussion after the improved version of the classified index has been typed. It may even prove necessary to go through this process several times, especially if different groups of experts are involved, as suggested in Section F0.3.3.

A special difficulty arises in these discussions from the fact that the classified index displays only part of the full classificatory structure, omitting cross-references to additional Broader or Narrower Terms, indications of semantic factors, and cross-references to Related Terms. As a result, many questions are asked that could have been avoided by displaying the full structure. On the other hand, it is very inconvenient to display the full structure without having a notation, and the notation should be assigned only after the classified index has been discussed thoroughly and has undergone major modifications arising from these discussions. So we have a vicious circle. The circle may be broken by indicating clearly that only part of the structure is displayed and by submitting additional information taken from the working file during the discussions.

F4.5 Assign Notational Symbols

The result of Step F4.4 is a first draft of the classified index or schedule. Only minor revisions are to be expected in the steps to follow. Therefore, it is now possible to assign a notational symbol to every preferred term included in the classified index, especially to every descriptor. Notational symbols for very specific concepts that are included only in the working file should be assigned after the main part has been revised in step F5.1. (As has been shown in Section D1.3.4, it is very advisable to have a notation if a classified index is to be part of the user version of the thesaurus. Even if one does not plan for a classified index in the user version, a notation might be useful in the construction of the thesaurus.)

Although the classified index is typed without a notation, enough space should be left at the left margin to enter the notation later on. Besides the original or stencil, one or two copies should be available. These copies serve as working copies for the design of the notation. When the design is finished, the notational symbols can be entered on the original or on the stencil, and the necessary number of copies of the finished draft of the classified index, together with notational symbols, can be made. See Section F5.9 for technical details.

F4.6 Make a Systematic Search for Additional Cross-References

The notations can now be used to record cross-references with less effort. Cross-references recorded sloppily before should now (or in Step F5.1(c)) be recorded precisely using notations. A systematic effort should be made to detect additional cross-references, which can then be recorded by their notations.

F5 COMPLETE FIRST DRAFT OF THE THESAURUS AS A WHOLE

F5.0 Introduction

The procedures to be used in the individual steps of this phase are very dependent on the size of the thesaurus and the technical means employed in thesaurus construction. This is particularly true for steps F5.2, "Produce main part"; F5.3, "Check inverse cross-references"; F5.4, "Duplicate preliminary version"; F5.6, "Enter modifications into master copy"; F5.7, "Making the alphabetical index"; F5.9, "Reproduce test version"; and F7 "Duplicate or print the final version". (Steps F5.1, "Revise working file" and F5.5, "Consultation with subject experts" are not dependent on the technical means employed.) Where differences exist, the procedures described in the

following are intended for smaller projects not using computer assistance, except possibly for the production of the alphabetical index.

F5.0.1 Special problems of smaller projects not using computer assistance (special topic)

The problem in this case is to avoid retyping the thesaurus over and over and to use at least parts of draft versions in the master copy for reproducing the final user version of the thesaurus. Therefore, the working file should not differ in its information from the main part of the user version. Accordingly, the less detailed cross-reference indicators should be used in this case, e.g., BT and not BT-WH (Broad Term-Whole). Only external spelling variants should be given in the working file; they will, of course, appear in the main part of the user version. (Internal spelling variants are needed only in computerized ISAR systems anyway. Because the size of the alphabetical index is small, external spelling variants can be used sparingly.) Furthermore, with small thesauri the working file on cards is needed only in thesaurus construction, not for updating. For updating a working copy of the user version is sufficient. On the other hand, if only very few people will be using the thesaurus, it may be possible to use the card file as main part so that retyping is not necessary at all.

The procedure described in the following is just one possibility. Alternative procedures are discussed in Section F9.

We have tacitly assumed that in smaller projects the user version of the main part will not be produced by typesetting. If, in fact, typesetting or complete retyping are envisioned, some of the restrictions mentioned disappear.

F5.1 Revise Entries in the Working File

Step F3 resulted in a preliminary main part in the form of the working file. In Step F4 much of the information in this file is disregarded in order to concentrate on the elaboration of the conceptual structure of the indexing language as represented in the classified index. We now come back to the working file in order to revise all the entries in the light of the results of Step F4. It is also appropriate at this stage to see to it that all terms conform to the rules selected for the form of entries (spelling, singular/plural, etc.). (If the cards for less important concepts have been removed from the working file in Step F3.3.6, they must be put back now at the appropriate location.)

For each card in the working file the following tasks must be performed (see the example in Figure 63):

(a) *Standardized abbreviations.* In many thesauri it is useful to use a standardized abbreviation instead of the full text of a descriptor (or even of a preferred term that is not descriptor) whenever the descriptor is referred to (as described in Section E1.8.3). If such a procedure is to be followed, one has to

0 | 1 | 2 | 3 | 4 | 5 | (6) | | check hierarchical level | 02 check type: (OS) OP NP EL CH | 03 Subject Field: M

05 Notation: M48.7 | 10 Color TV picture tubes(TH:659.5) (AR;SK)

12 Stand. abbr. (AB): M48.7 color TV pic tu

20 Spellings (incl. abbr.): Color television picture tubes (KL:TC904;SK)

30 Synonymous T. (ST) (incl. equiv. t.): Color kinescopes (TH) (AR); Color TV display tubes(TH)

4 Classification:

42 Category (CA):

44 Semantic factors/Broader Terms (BT): F78 Color TV receivers (TH) (AR;SK); (M48 TV picture tubes) ~~Kinescopes~~ (AR); ~~Color television set~~ (KL)

45 Is semantic factor of Narrower T. (NT) F13.7 Color TV screens (TH) (AR); (M48.7.3 Chromotrons)(AR); (M48.7.2 Shadow mask tubes)(AR;SK); ~~Lawrence tubes~~ (SK)

46 Related Terms (RT): (M48.2 Flat picture tubes(TH) (SK); Radiation hazards(TH) -V83)

50 Translations (TR):

F:

G:

R:

S:

60 Definition, scope note (SN): Descriptor introduced Jan. 1, 1960. In searching for documents indexed before that date, USE BT M48 TV picture tubes

70 Unspec. rel. (UN):

81 Editor/Date: AKS1959-07

Figure 63. Examples of revisions in the working file (F5.1). (The original card, as produced by merging, is shown in Figure 58 (Section F2.2). The top line and the notation have been filled in at earlier steps.)

formulate a standardized abbreviation for each descriptor and type a list of these standardized abbreviations. This task must be performed for the whole file first, before (c) can begin. (b) may be done together with (a) or together with (c).

(b) *Standardize form of Main Term.* The Main Term (the heading of the entry), as well as the terms listed in data fields ST and TRanslations, must be checked and, if necessary, changed so as to conform to the rules established for the form of terms (spelling, singular/plural, use of adjectives and verbs, etc.). If necessary, the appropriate changes should also be made in the classified index.

Depending on the use of the thesaurus it might be necessary to enter the unchanged forms of the term as external or internal spelling variants, as discussed in Section C6.2. Specifically, this is necessary in a cumulative thesaurus, as discussed in Section K1.3.

If one wants to avoid adjectives and/or verbs and if syntactical information is given for each term in the thesaurus, one may print out all entries that contain adjectives or verbs in any of the data fields MT, SP, ST, or TRanslations.

(c) *Standardize elements in BT, NT, RT.* All elements in the data fields BT, NT, RT must consist of notation (if any) and preferred term (or abbreviation of preferred term). If the notation is missing, it has to be inserted; if a non-preferred term is given, it has to be replaced by notation and preferred term. As a labor-saving device, one may put down just the notation and instruct the typist or the computer to fill in the term itself later. If the term appearing in the cross-reference cannot be found either in the classified index or the working file (blind cross-reference), a new card has to be made up and entered at an appropriate place in the hierarchy. (In big thesauri one should have a preliminary alphabetical index at this stage; this is possible if computer assistance is used.)

A somewhat tricky point arises here: While replacing non-preferred terms by preferred terms, one may detect that hierarchical relationships taken from different sources use different terms, but are the same conceptually.

Example:

On the card we have (sources 1 and 2)
Color TV picture tubes (1;2)
NT Three-gun color picture tubes (1)
Shadow mask tubes (2)
Since Three-gun color picture tubes is synonymous to the preferred term Shadow mask tubes, this reduces to
Color TV picture tubes (1;2)
NT Shadow mask tubes (1;2)
One of the original entries might in fact have been
Color Kinescopes (2)
NT Shadow mask tubes
Color Kinescopes being synonymous to the preferred term Color TV picture tubes. (Special rules hold for cumulative thesauri; see Section K1.3, esp. K1.3.1,1.3.)

All Broader Terms, Narrower Terms, and Related Terms that can be seen from

the classified index are now tagged so that they do not appear in the main part of the user version of the thesaurus unless they are part of a USE instruction. An example from Figure 22 (Section D0) illustrates this point.

Example:

Vacuum tubes
BT Vacuum devices
(Electron tubes)

Looking up Vacuum tubes in the classified index one can see there that Electron tubes is a Broader Term; therefore, Electron tubes is enclosed in parentheses on the card. (In the TEST model, all BT, NT, and RT are listed in the main part; therefore, no tags are needed.)

(d) *Improve classificatory structure.* Data field UN (Unspecified relationship) contains terms for which the proper type of relationship has not been determined previously; this determination should be made now. All BT, NT, and RT relationships should be checked as to their validity. The decomposition into semantic factors in particular has to be checked in view of the changes that have been made in the classified index in Step F4. This is also the time to check and/or fill in the information in the data fields 41 Subject field and 42 Facet (cf. Figure 21, Section C7).

(e) Enter *USE, SEE, and PT (Post to) instructions* and UF (Used for), SF (Seen from), and PF (Posted from) statements. If the crude lead-in form is used, this might involve some reshuffling of the terms on the card.

(f) *Create inverse cross-references and enter in the appropriate places.* Since these inverse cross-references are going to be checked in Step F5.3 anyway, the amount of care devoted to the task here should not be excessive. If, on the other hand, one were to delay all inverse cross-references until Step F5.3, a great many modifications would have to be made in the main part, necessitating extensive retyping. In cases illustrated by the following example cross-references can be limited (compare Figure 22a2, Section D0).

Example:

The descriptor M48.7 Color TV picture tubes has many narrower preferred terms that are not descriptors and that are listed after the descriptor in the main part. Each of them has an instruction

USE BT M48.7 Color TV picture tubes

Instead of entering all inverse UF NT statements, we may simply write

UF NT see the following nondescriptor entries*.*

The decisions made in this step may, in turn, give rise to modifications in the classified index. In principle, the situation is the same as discussed earlier: In step F3.3 the decomposition of a concept into semantic factors may give rise to the definition of a new concept, serving as a semantic factor. In the step here a concept may be used as a semantic factor in a new context, changing its definition or its place in the hierarchy. It is therefore important

to list for a concept A all the compound concepts that contain A as a semantic factor. This is achieved by entering all inverse cross-references.

Despite the fact that Steps F3.3 and F5.1 are similar in principle, there is a major practical difference: In Step F3 the emphasis is on the "distillation" of the classified index, i.e., the basic structure of the indexing language out of the wealth of the material collected; the expression of concepts not to be included in the indexing language is a by-product. In Step F5.1 it is the other way around: the emphasis is on the expression of concepts not included in the indexing language by concepts included, that is, descriptors; modifications in the list of descriptors to be included in the indexing language that may arise as a result of the work in this step are a by-product. (However, the systematic introduction of relationships among descriptors is part of Step F5.1).

F5.2 Produce the Main Part of the Thesaurus in List Form

Most of the cards in the working file are now likely to be messy due to many handwritten additions and modifications. It is therefore necessary to retype the main part of the thesaurus before the later steps can be performed. Steps F5.3, "Check inverse cross-references", F5.5, "Consultation with subject experts", and F6, "Testing of the thesaurus" are much easier if the main part of the thesaurus is available in list form, every entry being compressed to its actual size (not spread over an entire thesaurus form). Also, unchanged parts of this draft can be used in reproducing the final version of the thesaurus.

The best procedure is probably to simply type the main part in list form. An alternative is to type on blank cards which can then be shingled to produce a list copy. For technical details see Section F5.10.2.

The arrangement to be followed in typing the main part is described in Section D1.1.2 (assuming the Roget-Soergel-model): All Main Terms start at the left margin. Descriptors (marked by '+' after the hierarchical level and by crossing off DS in the top line of the thesaurus form) are emphasized by a solid underline. Preferred terms not used as descriptors (marked by '—' after the hierarchical level and by crossing off OP) are indicated by a broken underline. The hierarchical level is always given at the left of the margin proper; it is copied from the top of the thesaurus form, including '+' for descriptors and '—' for preferred terms not used as descriptors. BT, NT, and RT that are tagged "not for user version of main part" are omitted.

F5.3 Check Inverse Cross-References and Insert Where Necessary

It is important that for every cross-reference the appropriate inverse cross-reference is included, as discussed in Section C7.1. Checking the completeness of inverse cross-references involves a great many look-up processes.

It is therefore advisable to perform this step only after the main part of the thesaurus is available in list form, where speedy look-up is possible. In the Roget-Soergel model, only the classificatory structure (data fields BT, NT, RT) has to be considered in this step; the inverse cross-references for variants in spellings, for Synonymous and Equivalent Terms, and for translations are taken care of in the alphabetical index. In the TEST model *all* inverse cross-references must be checked. It is recommended that two persons together perform the somewhat cumbersome task of checking inverse cross-references. The additions that result are recorded in one of the working copies.

F5.4 Duplicate Preliminary Version of the Thesaurus

Enter the additions in the master copy leaving the picture as clean as possible (it might prove necessary to retype some of the entries; compare Section F5.10.2). Reproduce the classified index and the main part of the thesaurus in the required number of copies and distribute them among all the experts to be consulted in the next step.

F5.5 Review the Whole Thesaurus. Consult with Subject Experts

The synonym-homonym and the equivalence structure, as well as the classificatory structure, as worked out in the steps F3.3 and F5.1, are now presented in a form easy to peruse. In the next step the decisions made must be checked in consultation with experts. The following procedure is recommended; this procedure may be shortened if time does not allow for the full procedure.

(a) Discuss the different subject fields or subfields with a subject expert, entry by entry. The points to be considered are essentially the same as in step F4.4. The decomposition of compound concepts into semantic factors should be checked with special care. It is of advantage if one can repeat the same procedure with a second subject expert to get different points of view. Alternatively or concurrently, one may gather written comments from a number of subject experts according to F0.2.1(2). (It may be preferable to postpone the collection of written comments, which are not likely to be sent in promptly, to Step F6).

(b) Inform the subject experts in a field about the resulting modifications in that field.

(c) Probably there will remain a small number of problems that need further discussion (compare F0.2.3(3)): perhaps there has been disagreement between two subject experts consulted or between different written comments or a problem has been suggested for further discussion by subject experts after receiving information on modifications in (b). A special meeting has to be called to decide about these problems, as described in Section F0.2.3(b).

Instead of distributing the thesaurus to all subject experts involved after Step F5.4 and informing them afterwards about the modifications made in individual discussions, one may choose the following procedure: after F5.4, copies of the thesaurus are distributed only to the experts that are to be involved in individual discussions. After this step, the master copy is revised, and the number of copies necessary for the discussion in a meeting is produced and distributed. This procedure is more convenient for the subject experts involved; it has two disadvantages, however:

(1) it does not allow the collection of written comments in F5.5(a);

(2) it does not alert the subject experts—as does the information on modifications—to specific problems that should be considered more carefully.

F5.6 Enter Modifications in the Master Copy

The modifications and additions resulting from F5.5 are again entered in the master copy. In the same step inverse cross-references for newly introduced cross-references and the deletion of inverse cross-references for deleted cross-references are checked.

F5.7 Production of the Alphabetical Index (Technical)

By far the simplest method is to produce a KWIC (Key-word-in-context) or KWOC (Key-word-out-of-context) index by computer. Since computers and the appropriate programs are easily available nowadays, this method is described first. The programs available are usually written with titles and document numbers in mind. But they work just the same way on terms (instead of titles) and notations (instead of document numbers). The following description is oriented mainly toward the Roget-Soergel model.

F5.7.1 Production of a KWIC index

Process each card contained in the working file. Punch a separate card for each of the following: the Main Term appearing in data field MT; the spelling variants appearing in data field SP; every Synonymous or Equivalent Term appearing in data field ST; every translation appearing in data field TR, and all terms that appear in a UF (Used For) statement in the data fields NT or RT if separate main part entries for these terms are not made (intermediate form of lead-in). In a type-1 (lead-in only) multilingual thesaurus, one might have a separate alphabetical index for each language (compare Section D5.1). In this case a special code has to be punched for each language, e.g., F for French, G for German, so that they can be sorted into separate alphabets.

Punch the notation on each of these cards. If the program used has an option of punching the notation (in the program description: document number) first, this is preferable because this is the sequence used in the main part. A few rules have to be followed in this process:

(1) Split composite words by inserting a hyphen or a blank so that the second component shows up in the index, too. In some cases it might even be useful to separate prefixes.

Examples:

Gold-fish
Pre-test
Over-compensation

Make sure that a hyphen is considered as a separator between two words in the KWIC program used. One should also be aware of the possibility that prefixes separated from the word stem may appear in a stop-word list; this would mean that the full term starting with the prefix would not appear in the index. If the program used allows for using a non-printing separator instead of the hyphen this is greatly preferred.

(2) If a term is longer than the space provided on one punched card and if the program does not provide for continuation cards, break the term down into several KWIC lines. The omission of parts of the term in a KWIC line is indicated by '. . .'.

Example:

Term:	*B335 Military installations strengthening the offense potential*
Card 1	*B335 Military installations strengthening the . . .*
Card 2	*B335 . . . offense potential*
	or, even better,
Card 2	*B335 . . . strengthening the offense potential*

(3) If a notation consists of more characters than are provided in the program for the document number, truncate it to the required length. The truncated notation will still show where to look for the term in the classified index or the classified main part. The last sign of the truncated notation is immediately followed by a number sign in order to indicate that the notation has been truncated.

Remark: Procedure for including page numbers along with the notations. According to Section D1.5 the number of the page where the term appears in the main part in addition to the notation is not particularly useful. If page numbers are wanted nevertheless, they should be introduced only for the final version of the thesaurus in step F7. For purposes of the KWIC index the page number must form part of the punched card field provided for the document numbers. The easiest way to introduce them is as follows: Leave the appropriate columns blank while punching the cards. If all the modifications introduced in step F6 have been inserted into the punched card file and if the main part of the thesaurus is available in its final form with page numbers, insert page number cards into the punched card file at the appropriate places. (Remember that the punched cards serving as input for the KWIC program are in the sequence in which the terms appear in the main part.) Gang-punch the page numbers into the appropriate packs of cards, take the

page number cards out, and the punched card deck is ready for running the KWIC program. In this case, continuation cards must not be used even if the KWIC program does provide for them, and the procedure described in (2) has to be followed.

F5.7.2 Manual production of the alphabetical index

For the manual production of the alphabetical index, three steps are necessary:

(1) Produce the necessary entries;
(2) sort into alphabetical order;
(3) type the index.

The main problems arise with step (1), producing the necessary entries. There must be one entry for each form in which a term should appear in the alphabetical index, e.g., one entry for the direct form and one for the inverted form of a term (this problem is avoided in a KWIC index). Each entry must contain the information to be given in the alphabetical index, namely, the notation and/or the text of the preferred term and/or the page number of the preferred term in the main part. Sometimes it is convenient to produce the entries for the alphabetical index by copying; in this case they often contain more information than is needed. The typist must be instructed properly what information to type. In producing the entries for the alphabetical index one starts, as in F5.7.1, from the entry for the preferred term. One possibility is to write an index card or a paper slip for each entry for the alphabetical index, that is, for the Main Term itself, possibly for different forms of the Main Term (such as inverted form), for each spelling variant, and for each Synonymous or Equivalent Term (and possibly for different forms of each of these terms). The other needed information as described above is added. Another possibility is to produce a number of copies of the entry for the preferred term and underline on each copy a different spelling variant or Synonymous or Equivalent Term. The underlined term is then the entry term for the alphabetical index. In this case inverted forms have to be entered on the card; the most appropriate field is SP-EX (external spelling variants). If one uses translucent paper for the thesaurus forms, the necessary copies can be produced by diazo copying.

F5.7.3 TEST model: produce alphabetical main part and alphabetical index

As was shown in the beginning of Section F3, the development of any thesaurus should be based on a working file in classified arrangement. In the TEST model, the user version of the main part is alphabetical. To produce it, the necessary entries are created and then sorted in alphabetical order. All cards in the working file are entries. Further entries are created from the

cards in the working file, as illustrated by the following examples (taken from Figure 18, Section C5.1).

Examples:

(1) Crude lead-in form

Card in working file

Data processing

UF Automatic data processing

Electronic data processing

Data analysis

Data management

Data handling

BT . . .

Additional entries created

Automatic data processing

USE Data processing

. . .

Data handling

USE Data processing

(2) Detailed lead-in form

Cards in working file

Data processing

UF ST Automatic data processing

Electronic data processing

UF NT Data analysis

Data management

BT . . .

Data analysis

USE BT Data processing

Data management

SF ST Data handling

USE BT Data processing

SN . . .

Additional entries created

Automatic data processing

USE ST Data processing

Electronic data processing

USE ST Data processing

Data handling

SEE ST Data management

USE BT Data processing

As can be seen from these examples, elements in UF NT are not processed

to result in a USE BT instruction. However, if the intermediate form of lead-in is used, the working file will not contain a card for, e.g., *Data analysis,* and the elements in UF NT must be processed to obtain the proper USE BT instruction.

In both cases entries should be made for external spelling variants.

Example:

Aesthetics
USE Esthetics.

The alphabetical index is best produced as a KWIC or KWOC index of the Main Terms in all the main part entries (original and created). The notation is simply omitted if the thesaurus being built does not give a notation. The alphabetical index might also be produced as described in Section F5.7.2. However, in many cases the proper form of the term would have to be added in a USE instruction.

Example:

Antennas, radar *USE Radar antennas.*

Another possibility is to include all forms of a term as entries in the main part. (This is done in the Library of Congress Subject Headings; see Section D1.7.10.)

F5.7.4 Remark

Some authors suggest establishing two card files from the very beginning of the construction of a thesaurus—one classified and one alphabetic—and the production of two cards for each of the collected terms. If the alphabetical file were to have any meaning for the production of the alphabetical index, this would mean that one would have to enter in the alphabetical file all the decisions made during thesaurus development; that obviously is not possible in terms of economics. It would be useful to have an alphabetical file during thesaurus construction so that one could see whether or not a particular term is already included in the thesaurus and to which subject field or subfield it belongs. Without the tool of an alphabetical file, one has to rely on his memory for such inquiries. But even that doesn't justify the effort of keeping a separate alphabetical file. However, if computer-assistance is used, a preliminary alphabetical index can be printed at various stages of thesaurus construction. This is very useful, especially for big thesauri.

F5.8 Check Homonyms and Improve Cross-Reference Structure Using the Alphabetical Index

If the same term has been assigned two different notations or otherwise occurs twice as preferred term or if a term has been used as a synonym of two

different preferred terms, this shows up in the alphabetical index. Either the term is homonymous and should be disambiguated, or something is wrong in the synonym-homonym structure or in the equivalence structure. In a KWIC index it will also become apparent if the same word occurs with different meanings in different multiword terms.

Also the alphabetical index collocates terms that contain common or similar words. Often this indicates that there are conceptual relationships between these terms, and appropriate cross-references should be made in the thesaurus. Thus, to the extent that conceptual structure is reflected in the linguistic structure of terms the alphabetical index may be used to improve the cross-reference structure. The suggested format of the alphabetical index offers particular advantages in this connection: if two terms that appear near to each other in the alphabetical index are also collocated in the classified sequence (as may be seen from the notation), no action is necessary. If they are not collocated in the classified sequence, then one can look up one of the terms in the main part using the notation and check whether an appropriate cross-reference has been introduced. If not, it has to be decided whether it is a hierarchical or a Related Term cross-reference. One should take care to record the inverse cross-reference, too (unless an additional run using a computer program for this purpose can be made).

F5.9 Reproduce Test Version of the Thesaurus

The following parts of the thesaurus are now ready:

—classified index;
—main part;
—alphabetical index.

If one wishes to have additional displays (such as an overview of the subject fields and subfields, a display of the checklist descriptors, or graphical displays) they can be produced now. The entire thesaurus can then be reproduced in the number of copies which is required for the practical test to be performed in Step F6.

F5.10 Remarks on Some Technical Problems Arising in F5, F6, and F7 (Technical)

F5.10.1 Use of notations as "shorthand" for descriptors

Wherever a descriptor has to be entered in the process of modification, it is sufficient to give its notation. The term itself (or a standardized abbreviation) can be added later on by the typist. For this purpose the typist is provided with a listing that gives for every notation the appropriate term or standardized abbreviation. This listing is compressed into the smallest space possible to minimize page-turning.

F5.10.2 Technical considerations as to the production of the main part of the thesaurus in smaller projects without computer assistance

Since a considerable amount of clerical work goes into the production of the main part of the thesaurus, it is worthwhile considering in more detail some seemingly trivial questions connected therewith. It is best to start by listing the requirements to be met:

—Two working copies of the main part are needed for step F5.3, "Check inverse cross-references".

—A number of working copies of the main part, as modified by the results of step F5.3 "Check inverse cross-references", are needed for step F5.5, "Consultation with subject experts".

—A number of working copies of the main part, as modified by the results of F5.5, are needed in F6, "Testing the thesaurus".

—The main part, as modified by the results of F6, has to be duplicated or printed in step F7.

—The final format should be about letter size, two columns per page. Reduction by 1:1.4 should be used in preparing the final copy (linear reduction 1:1.4 makes for a reduction in area of 1:2).

A good solution for meeting these requirements is the following: In step F5.2, the first typing of the main part, a master is created that has the following characteristics:

(a) It is easy to correct, even after copies have been made.

(b) It is possible to make copies repeatedly from it (that is, the master can be stored after copies have been made).

(c) It is possible to apply photographic reproduction processes involving reduction in size.

The only materials showing these qualities are normal paper, which may be used as a master for Xerox copies or other photographic reproduction processes, and translucent paper, which can also be used for diazo copying.

In typing, a column width of 6 inches (including margins) should be used. This allows for appropriate reduction in size in reproducing the final version (12:1.4 = 8.5, assuming letter size for final format). It also leaves a wide margin which can be used for insertion of comments that should not show up in the final version but that are needed in the draft version for discussion, etc.

As to further procedure, one may follow one of two alternatives:

(1) In step F5.2, type the main part on sheets of paper. Modifications from checking the inverse cross-references (Step F5.3) and consultation with subject experts (Step F5.5) may be entered on the margin or, if this becomes too messy, the whole entry is retyped and pasted over the old entry (or the correct entries of the original master are cut and mounted together with the new entry on a new

sheet). If many entries on a page have to be typed, it is best to retype the whole page. The master for producing the final version is prepared as follows: Cut the entries to a width of 6 inches (including margins) and mount them in two columns on 12″ × 15.4″ sheets. These sheets are then ready for photographic reproduction with a reduction of 1:1.4. Of course, all entries having modifications indicated on the margin must be retyped before this step.

In this procedure the two working copies needed for checking inverse cross-references (Step F5.3) can be produced as carbon copies while typing the main part in Step F5.2. (If translucent paper is used as a master, the copies needed for the consultation with experts (Step F5.5) and for the test phase (Step F6) may be produced by diazo copying, which might be cheaper than other processes. However, this process is not compatible with the technique of mounting retyped entries described above.)

This procedure also has a major disadvantage; the replacement of entries by the technique of mounting is cumbersome. If half of the entries on a page have to be retyped, it is usually easier to retype the whole page, thus unnecessarily duplicating the typing of the correct entries.

(2) The alternative is to type each entry on a separate card. This procedure offers more flexibility in the replacement of entries. Blank cards (not thesaurus forms) are used, and the entries are typed in compact form as they should appear in the list. The master, then, is a card file. For making copies the cards are shingled so as to show the typed part only. This procedure is definitely to be preferred if the main part is to be arranged alphabetically; the cards can easily be sorted in alphabetical order.

With this procedure carbon copies and use of translucent paper for diazo copying are not feasible. In addition, the making of copies is cumbersome unless specific devices for mounting the cards are available.

F6 TEST THE THESAURUS BY INDEXING AND RETRIEVAL EXPERIMENTS

Index 1,000 to 2,000 documents with the aid of the thesaurus. In addition, collect as many potential search requests as possible (or use the search requests collected in the collection phase) and formulate these search requests with the aid of the thesaurus. Perform a number of searches in the test collection and analyze the search results with respect to search failures due to shortcomings of the thesaurus. In analyzing the search results one should keep in mind that they are dependent on many factors other than the quality of the thesaurus. The most important of these factors is the selection of documents in the test collection and the quality of indexing (which depends only in part on the quality of the thesaurus). (This is not the place to go into details of the evaluation of ISAR systems; the reader may consult the references given in the back.)

If, for reasons of time, it was not possible to collect written comments

from numerous subject experts in step F5.5, "Consultation with subject experts", such a collection can be done now.

As a result of this step, new terms are entered both in the indexing language and in the lead-in vocabulary, definitions are broadened or narrowed down, new relationships between concepts are detected. All these additions and modifications are entered into the working file, the user version of the main part, the classified index, and the alphabetical index, as appropriate. (See Chapter J on thesaurus updating for procedural details.) The thesaurus is now complete. This does not mean, of course, that the thesaurus is perfect and that no further improvements and modifications are possible. On the contrary, the thesaurus has to be updated on a continuing basis, as discussed in Chapter J.

F7 DUPLICATE OR PRINT THE USER VERSION OF THE THESAURUS

F7.1 Duplication or Printing of the Main Part and the Alphabetical Index

With the main part and the alphabetical index, there are no specific problems as to layout and space limits. If the detailed scheme of cross-reference indicators has been used in the working file, one should make sure in this step that all specific cross-reference indicators are replaced by the corresponding general ones; for example, BT-CL (Broader Term - Class inclusion) is replaced by BT. Also, the working file contains BT, NT, and RT cross-references that can be seen easily from the classified index. These cross-references are tagged as "not to be included in the user version", and they are omitted accordingly. (Compare Section F4.0, "Classified index and cross-references in BT, NT, and RT in the main part".) In smaller projects projects not using computer assistance, these problems are already taken care of, as discussed in Section F5.0.1.

F7.2 Duplication or Printing of the Classified Index

The classified index, especially the list of checklist descriptors, has to be presented in such a way as to make its perusal as easy as possible. Different type fonts and other means may be used to achieve this purpose. It is often useful to introduce the requirement that the display of checklist descriptors should fit onto a sheet of double letter size. This offers the advantage that the checklist descriptors can be surveyed with one glance. This requirement becomes almost mandatory if one wants to print the checklist descriptors on the document or request analysis sheet. If one introduces this requirement, one may have to reconsider the selection of some of the checklist descriptors for reasons of space. It is recommended, therefore, that the typesetting be

done before Step F6 (use monotype rather than linotype because of the many corrections to be expected). This ensures that the list of checklist descriptors (which in some systems is identical with the indexing language) can be printed in the space available. Due to many cross-references involving notational symbols and to the usage of a number of different type fonts, the typesetting is very complicated. One should therefore weigh the possibility of letting the type stand so that later revised versions can be printed.

F7.3 Proofreading

It will be necessary to be especially thorough in proofreading the thesaurus.

F8 FURTHER REMARKS CONCERNING THE WORK-FLOW AND MODIFICATIONS OF THE STANDARD WORK-FLOW

F8.0 Introduction

The optimal sequence of steps in thesaurus construction is dependent on so many factors that it has to be determined for every individual project. The sequence shown in Figure 53, Section F0.1, should be used only as a guideline, and appropriate modifications should be made as the individual case requires. The following considerations should be helpful in these decisions. It is recommended that the reader look at the flowchart in Figure 53 while going over the following discussion.

F8.1 Sequence of the Steps F3, "Work Out the Preliminary Structure of the Thesaurus" and F4, "Work Out the First Draft of the Classified Index"

One possibility is to complete step F3, "Work out the preliminary structure of the thesaurus" for all the subject fields and then move to Step F4, "Work out the first draft of the classified index". This has the following advantage: In working on subject field 3, for example, one may detect by semantic factoring concepts which belong to subject field 1 or 2. These additional concepts are then included before step F4 is performed for these subject fields. On the other hand, there is an inverse problem: in working on subject field 3, it might be necessary to use descriptors from subject field 1 as semantic factors. If the area of the thesaurus is so large that the descriptors of the previous subject fields cannot be kept in mind, this is difficult. In this case it is recommended that step F4.1 or even all steps F4.1 through F4.4 be completed for a subject field or group of subject fields before step F3 is performed for the next subject field. Then the classified index for the prior subject fields can be used to look up the descriptors needed.

Another viewpoint for the planning of the steps F3 and F4 is to make optimal use of the staff and distribute the workload evenly over time. If, for

example, step F3 is completed for all subject fields before step F4 is performed, then the typist may be idle for that time and afterwards in a big rush to type the classified index.

F8.2 When Should the Notation Be Introduced?

In the standard sequence the notation is introduced after the classified index has been discussed with experts and is somewhat "stabilized". If one does not expect many modifications from the discussion with experts in Step F4.4, then it is useful to introduce the notation before this step. This would make the discussion easier because in the discussion descriptors could be cited by their notation and it would be easier to locate the descriptors in the classified arrangement.

On the other hand, it is possible to postpone the assignment of a notation until the modifications in Step F5, "Prepare first draft of the thesaurus as a whole" have been completed. Advantages of introducing the notation before Step F5 are as follows: Many procedures in F5, as well as Step F6, "Testing the thesaurus through indexing and retrieval experiments", are much easier if a notation is at hand. In Step F5.1, "Revise entries in the working file", descriptors or other preferred terms appearing in cross-references need not be written in full but can be cited by their notation. The same is true for the checking of inverse cross-references in Step F5.3 and the making of the alphabetical index in Step F5.7. The disadvantages of introducing the notation before Step F5 are as follows: In Step F5 the classificatory structure is modified, particularly in Step F5.1, "Revise entries in the working file". This means that the notations have to be modified accordingly. If the notation for a descriptor is changed, this change has to be recorded at every place where this descriptor appears in a cross-reference. This is less difficult than it might seem at first; after the check of inverse cross-references in F5.3, the entry for a descriptor gives all the places where this descriptor appears and where consequently the change in notation has to be recorded.

Of course, changes in notation necessary due to modifications of the classificatory structure in F5.1, "Revise entries in working file" are introduced before the first draft of the main part is typed in Step F5.2, and the same is done in the following steps. If one has decided not to introduce the notation before Step F5.1, one might introduce it either after Step F5.1, after Step F5.3, "Check inverse cross-references", after Step F5.5, "Consultation with subject experts", or even after Step F6, "Testing the thesaurus". The later the notation is introduced, the smaller are the advantages to be gained from the notation in thesaurus construction. At the same time, the effort for changes in notation is reduced.

F8.3 When Should the Main Part Be Typed (Smaller Projects without Computer Assistance)?

In the standard sequence, the main part or at least part of it has to be typed twice, namely, in step F5.2, "Type the main part of the thesaurus in list form" and in reproducing the user version (Step F7.1). This can be avoided if punched paper tape or punched card equipment or a computer are available. However, the following discussion deals with the case where such equipment is not available. Advantages and disadvantages of several possibilities will be considered. (It might be useful at this point to re-read Section F5.10.2.)

(1) It is not useful to type the main part before Step F5.1, "Revise entries in the working file". This step can very well be performed using the cards in the working file, so nothing is gained by prior typing. On the other hand, numerous modifications arise in Step F5.1, and all these modifications would have to be inserted into the typed version.

(2) Type the main part before Step F5.3, "Check inverse cross-references". This has the following advantage: Checking the inverse cross-references in Step F5.3 is much easier (an estimated time-saving by a factor 2 to 4) if the main part is available in list form.

(3) Type the main part after the inverse cross-references have been checked in Step F5.3. Advantage: the copies of the main part used for the discussions with experts in Step F5.5 would be more orderly, since the modifications resulting from Step F5.3 would not appear as handwritten additions. On the other hand, one would lose the advantage mentioned in (2).

(4) Type after Step F5.5, "Consultation with subject experts", before Step F6, "Testing the thesaurus". This will rarely be advisable because of the following major disadvantages: The consultation with experts would have to be based on the working file on cards, which would be rather messy due to the many modifications entered during the process; the file would have to be duplicated in the necessary number of copies, and this might be just as expensive as typing. On the other hand, this procedure would have the following advantage: the draft used for testing through indexing and retrieval experiments in Step F6 would be very orderly; in fact, it would not differ too much from the final user version because the modifications suggested by the experimental indexing are usually limited (provided the initial collection of terms has been performed adequately). This procedure would therefore have the additional advantage of minimizing the number of entries in the main part that have to be retyped in producing the user version.

(5) Type after F6, "Testing the thesaurus". From the end of the previous paragraph, it follows that this is not advisable; one would lose the advantage of having an orderly draft for testing the thesaurus without saving much work in typing.

F8.4 Drawing Up and Using a "Core Classification" Consisting of Elemental Concepts Early in the Process

(Before going further it might be advisable to reread Section C2.5 on the idea of a core classification and possibly Section C2.8 on the optimization of an indexing language where the number of descriptors is limited).

Some systems require that the indexing language consist only of a limited number of elemental or nearly elemental concepts (to give a concrete example: the indexing language for a peek-a-boo system). In this case the indexing language is restricted to a "core classification". But even in a system using many precombined descriptors, a core classification of limited size is useful for achieving compatibility and/or for establishing an auxiliary ISAR system. To keep the core classification limited, it might be necessary to force a decomposition of some concepts into semantic factors that are available in the indexing language or the core classification as discussed in Section C2.8.2. In this case it is recommended that a tentative core classification be drawn up after Step F3.3.1, "Working out the synonym-homonym structure and the equivalent structure". In Step F3.3.2, "Working out the classificatory structure", which involves the decomposition of concepts into semantic factors, the tentative core classification gives some idea what elemental concepts are available as semantic factors. Of course, new elemental concepts may still be created. A consolidation will take place in Step F4, "Work out the first draft of the classified index".

One may take an even more radical approach: A tentative list of elemental concepts is defined right at the outset after looking at some of the sources. The decomposition of compound concepts into semantic factors, resulting in a combination of elemental concepts for each compound concept, can then be done *before* Step F3, "Work out the structure of the thesaurus". This opens the possibility of grouping the terms according to the concept combinations assigned to them. For example, two terms having exactly the same concept combination are very likely to be synonymous. Of course, this is checked in an editing step: if two terms are not synonymous and have been assigned the same concept combination in spite of this, one should consider adding descriptors to the indexing language or the core classification so that a distinction between the two terms is possible. The derivation of further relationships from the concept combinations is described in C1.3. One could say that this procedure operates as follows: A combination of elemental concepts is assigned to each term; this is a *local* operation. The *global* structure is then derived automatically or, at least, with computer assistance. We shall discuss in Section G3.4.2 how this approach can be implemented in computer-assisted thesaurus construction.

F8.5 Extending the Collection of Conceptual Relationships, Especially for Cooperative Information Services

If heavy emphasis is placed on the complete collection of conceptual relationships as seen from different points of view one may proceed as follows: After the source lists of terms have been collected, they are first presented to scientists who are asked to indicate Synonymous and Equivalent Terms, Broader and Narrower Terms, Related Terms, and possibly decomposition into semantic factors, using a preliminary core classification as described in the previous section; these enriched lists of terms are then processed as described in Section F2. This approach is particularly appropriate if the task is to develop an indexing language and thesaurus that is to serve as the basis for the cooperation of a number of information service institutions. If the enrichment is done at each of these institutions (each institution processing appropriate parts of the over-all input vocabulary), all viewpoints are brought out in the indexing language. This is essential for the success of cooperation, as will be discussed in detail in Chapter K. On the other hand, this procedure means increased effort in thesaurus construction, since most terms are analyzed by at least two people.

F9 USE OF PUNCHED PAPER TAPE AND PUNCHED CARDS IN THESAURUS CONSTRUCTION (SPECIAL TOPIC, IN PART TECHNICAL)

F9.1 Use of Punched-Paper-Tape Typewriters in Thesaurus Construction

The use of punched-paper-tape typewriters for more efficient text processing is well known. Of particular interest in our context is the use of punched-paper-tape cards. These are cards that are punched on one edge like punched paper tape; they could also be called "unitized punched paper tape". Punched-paper-tape cards are easier to sort manually and to correct than punched paper tape.

F9.1.1 Modifications in the flow of work

Only small modifications are necessary. It is not worthwhile to produce a punched paper tape in the initial stages of collection of terms even if terms from different sources are typed on thesaurus forms (rather than using cut-and-paste techniques). The reason is that during the process of thesaurus construction so many modifications are introduced that an initial punched paper tape would be of no value. The main savings can be achieved in the typing of the main part of the thesaurus, as discussed in Section F8.3: The main part can be typed after Step F5.1, at the same time producing a

punched paper tape, (or, even better, punched-paper-tape cards). In later steps a modified punched paper tape can be produced without retyping the correct parts, and a modified listing can be typed automatically using the modified tape.

A number of technical notes are of interest:

(1) One can make use of a suitable control tape to enter function codes into the punched paper tape in such a way that it is possible to write on thesaurus forms with the appropriate spacing and also on paper in compact list form. This makes it possible to obtain a clean copy of the card file without much effort. This card file is convenient for updating.

At the same time, information needed for the working file but not for the user version (such as spelling variants) can be tagged by function codes so that it does not appear when writing in list form.

(2) Punched-paper-tape cards are to be preferred because only those cards where modifications have to be made need to be fed into the machine and duplicated. If punched paper tape is used, appropriate function codes should be entered so that an entry can be duplicated at high speed without typing and the machine will stop automatically before the next entry. In this way correct entries can be duplicated into the modified tape much faster.

(3) If many modifications are necessary in an entry, it is faster to retype it in its entirety than to duplicate the correct parts and insert the modifications. The limit point depends on the ability of the typist.

(4) Word of caution: The production of a corrected duplicate tape and the automatic typing from punched paper tape takes more time and effort than many sales representatives might have you believe, especially if the equipment is low-speed (a good punched-paper-tape typewriter writes about 900 characters per minute from the tape) and/or if the equipment is not suitable for continuous high-speed operation.

F9.1.2 Conversion of punched paper tape to punched cards

Punched paper tape can be converted to punched cards (for example, by the IBM 47). These cards can then be used for the following purposes:

(1) Production of various listings using conventional punched-card equipment as described below.

(2) Data input into a computer if the computer program requires card input and/or if no punched tape reader is available at the computer installation used. This can be especially useful for producing cards as input for a KWIC program. The punched paper tape is formatted as follows: the notation is preceded by a special code (start of record code); the preferred term, the spelling variants, and the Synonymous and Equivalent Terms are preceded by another special code (field delimiter code); the tape-to-card converter can be programmed to store the notation and to make up a punched card for every term, the stored notation being punched in specified columns. Very long terms should be divided by the field delimiter in the punched paper tape so that two or more punched cards are produced.

F9.2 Use of Conventional Punched Card Equipment

F9.2.1 Punched-card-controlled typewriters (for example, the IBM 870 Document Writing System)

There are punched-card-controlled typewriters that can be used in the same way as punched-paper-tape typewriters. Advantages and disadvantages are the same as in other applications. Punched cards are easier to correct but more difficult to produce because of the 80-column limit. Punched cards are also more expensive.

F9.2.2 Keypunch and unit-record equipment

A combination of keypunch and accounting machine (tabulating machine) may be used in the same way as a punched-card-controlled typewriter. Additional disadvantages are:

—limited character set (in particular, no lower case letters);
—one-to-one correspondence of punched cards and lines in the printout (the types of programming to overcome this would not be practical in our context).

An advantage is the higher speed of accounting machines in printing.

Instead of a keypunch, a punched-paper-tape typewriter and subsequent conversion of the paper tape to punched cards can also be used.

A sorter in connection with an accounting machine can be used for the production of the alphabetical index or other listings in specified order if the appropriate sort-key is punched in the cards. In the alphabetical index it would not be practical to give more information than the notation for each entry, since each entry consists of only one punched card. Inverted forms of the preferred term, as well as Equivalent or Synonymous Terms, must be entered as spelling variants in order to appear in the alphabetical index. A collator could be used to detect duplicates.

An interesting variation of the procedures described is the following: In the working file, use punched cards as thesaurus forms on which the terms and other information are transferred. Then punch part of the information. Gang-punch the source code. The cards can now be sorted, according to a subject field and subfield code, and a lexicographer can work on the cards so sorted. The cards can also be sorted by other sort-keys. The problem with this approach is that the writing space on a punched card is limited and only part of the data is machine-readable. The production of an alphabetical index, for example, would need additional punching.

(Edge-notched cards can be used in a similar way; however, the sorting operations are much slower and listings cannot be produced automatically.)