P.J. Whitelock
University of Manchester
Institute of Science and Technology, England

A Descriptor Bank of Social Science Terms

Whitelock, P.J.: A descriptor bank of social science terms.
In: Int. Classif. 9 (1982) No. 3, p. 145–151, 14 refs.

This article presents proposals for the organization and creation of a Descriptor Bank of terms used in Social Science Thesauri. The introduction places the present study in its context as one stage of a long-term project directed toward the creation of an Integrated Thesaurus for the Social Sciences. Subsequent sections describe the content and layout of Descriptor Bank entries, the record formats used in existing thesauri, and the software necessary for the creation of the Descriptor Bank and its subsequent use as a tool in the development of the Integrated Thesaurus. Finally, estimates of the human and machine resources necessary are presented. (Author)

0. Introduction

This article presents the result of a project entitled 'A Feasibility Study of the Creation of a Descriptor Bank of Social Science Terms', carried out by the author at the Centre for Computational Linguistics at UMIST, Manchester, for the Division for the International Development of the Social Sciences of UNESCO.

The present study is a part of a long term project whose objective is the creation of a Integrated Multilingual Thesaurus for the Social Sciences. This project was formally initiated by UNESCO at a Consultative Meeting in June 1980 (14). This meeting also proposed a series of short term projects whose successful completion would constitute the essential groundwork for the establishment of the Integrated Thesaurus. The short term projects of the first stage were: the compilation of a bibliography of dictionaries, thesauri, and classification systems in the social sciences (7); the development of guidelines for the establishment of comparison and compatability matrices between thesauri (10, 11, 12), and between classification systems (4, 5); preliminary testing of the guidelines using several small subsets of the relevant data (1); and finally the Descriptor Bank feasibility study presented here. With the completion of these projects, the development of the Integrated Thesaurus will proceed to the second stage, the establishment of the Descriptor Bank.

The Descriptor Bank was first proposed by J. Meyriat in a paper presented to the Consultative Meeting mentioned above (8). After a survey of existing documentary languages (bibliographies, classification schemes and thesauri), he concluded:

'none of the existing documentary languages offers a firm ground for the building of a General Indexing Language (GIL) for the whole social sciences . . . As a basis for the preparation of such a GIL, it would be advisable first to select a limited number of existing languages . . . and then, to compile together the descriptors contained in each of these languages, care being given to preserving all hierarchical and other relations of each descriptor.'

The principal function of the Descriptor Bank is to serve as a focus for the collection of information from many divergent sources. By converting this information to a common format, thus eliminating trivial variations in its representation, important differences in the content and organization of the incorporated thesauri become more easily discernible. Information available in this form is not only essential as a reference for the compilers of the Integrated Thesaurus, but also fulfils other useful functions. It can be used as a switching tool between different Indexing Languages when several thesauri are being accessed together. It will also serve as a master reference for a variety of terminological activities, as a translation aid, and as a source of peripheral terms for compilers of specialised thesauri. In the long term, these latter functions may be taken over by the Integrated Thesaurus.

1. Organization of the Descriptor Bank

The essential organization of the Descriptor Bank will be alphabetical by term (descriptor and lead-in term). Such an ordering is defined, or definable, on all candidate thesauri, albeit with variations (see 1.3.3). In contrast, a common classified ordering does not exist: it is, of course, the ultimate function of the Integrated Thesaurus to impose one. In fact, the majority of thesauri do not provide a unique class-mark for individual indexing terms. However, inclusion in the alphabetical Descriptor Bank of such classification information as is available presents no problems, and this may be useful in the definition of certain output displays (see 3.5).

1.1 Attributes of the Descriptor Bank

The Descriptor Bank will provide the compilers of the Integrated Thesaurus with a large volume of information enabling them to make decisions about the terms and relations between terms to be incorporated in the Integrated Thesaurus. Brief descriptions of some of these attributes are given below. (See Aitchison (1, 2) for more extensive coverage with examples).

1.1.1 Descriptors and Lead-in Terms

The Descriptor Bank shows which descriptors and lead-in terms occur in which thesauri, and which terms occur as both in different thesauri.

1.1.2 Hierarchical Relationships

The Descriptor Bank provides a comprehensive representation of the hierarchical relationships of each descriptor. For instance, it shows which thesauri have many broader, narrower and related terms, and which sub-terms have different relationships to the main term in different thesauri (e.g. term X is NT to term Y in thesaurus A, but RT to term Y in thesaurus B).

1.1.3 Foreign Language Equivalents

The Descriptor Bank will give foreign language equivalents of terms, and will highlight particular incompatibilities, such as one term being rendered differently into French in different thesauri.

1.1.4 Definitions and Scope Notes

The inclusion of definitions and scope notes will assist in the recognition of those cases where a single term has two or more meanings (partly or wholly unrelated) in different thesauri (polysemes and/or homographs).

1.2 Types of Descriptor Bank Entry and their Output Layouts

It is proposed that the principal output layout of the Descriptor Bank should comprise two parts for each entry:

i) The logical union of the information associated with the entry term in all thesauri, formatted in a similar fashion to the entries in a conventional alphabetical thesaurus listing.

ii) A matrix of binary (presence/absence) indicators placed alongside part i, with a column for each constituent thesaurus and a row for each line of the entry. The matrix elements themselves would most usefully be printed as a two letter mnemonic code, which would be the same as the column heading. However, these thesaurus codes could be reduced to a single character (e.g. X) indicating presence, were this neccessary in order to fit all columns of the matrix onto a single output sheet. For added ease of reference to the Descriptor Bank printout, it is suggested that the line type indicator (e.g. Main, BT, UF etc., or corresponding international symbol) is repeated at the right hand end of the matrix row.

1.2.1 Descriptor layout

Example:

	AP	*IC*	*IL*	*IS*	*MA*	*NC*	*PM*	*SP*	*SS*	*UT*	
NUPTIALITY		IC	IL	IS	MA		PM	SP		UT	MAIN
R10.71										UT	CLASS
13.01.00							PM				CLASS
14.01			IL								CLASS
14.02.05					MA						CLASS
15421				IS							CLASS
22230		IC									CLASS
F= Nuptialite			IL	IS	MA		PM				F=
S= Nupcialidad			IL		MA		PM				S=
= Marriage rate								SP			=
< Population Dynamics								SP			<
< Population Event										UT	<
– Divorce								SP			–
– Families								SP			–
– Family System								SP			–
– Genetic counselling								SP			–
– Marriage		IC	IL	IS	MA		PM				–
– Nuptiality rate							PM				–
– Nuptiality table							PM				–
– Sexual union							PM				–
– Statistical data								SP			–

Note the variation in the layout of main term and class codes from that proposed by Aitchison, where they were in the form:

NUPTIALITY
- IC : 22230
- IL : 14.01
- IS : 15421
- MA : 14.02.05
- PM : 13.01.00
- UT : R10.71

with no associated matrix. This is not a significant difference, but it is felt that the form proposed herein is more revealing, and it will simplify the display software.

1.2.2 Lead-in Term Layout

A term which occurs as a lead-in term in one or more thesauri should be assigned to a separate Descriptor Bank record, even if it also occurs as a descriptor in another thesaurus. If this is not done, it will not be possible to tell the status of the term in a given thesaurus from the matrix row associated with the term itself. In the alphabetical listing of the Descriptor Bank, the entries for the term as descriptor and lead-in term, if both exist, will be adjacent, and may be visually compared with ease.

Layout can be according to standard thesaurus practice, with the assosiated matrix as in descriptor layout, e.g.

	AP	*IC*	*IL*	*IS*	*MA*	*NC*	*PM*	*SP*	*SS*	*UT*	
X		IC					PM				Lead
→ Y		IC									→
→ Z							PM				→

This shows that term X occurs as a lead-in term in both IC and PM, but that in the former, Y must be used as the descriptor, and in the latter, Z.

1.2.3 Layout of Combined Descriptor and Factored Lead-in Term

In some thesauri, two or more descriptors are used in place of a compound lead-in term (factoring). This is usually shown thus:

a) Lead-in Term
   → Descriptor
   \+ Descriptor

b) Descriptor
   \+ Descriptor
   UF Lead-in Term

This layout, with matrix, is suitable for use in the Descriptor Bank.

Records of both types must be provided. There will be an entry of type b for each term in a compound, and a set of entries of types a and b for each pattern of compounding (e.g. X + Y, X + Z, X + Y + Z).

In the remainder of the report, when the type of a record (as opposed to the record format type, or field type) is discussed, it refers to the categories described in this sub-section.

1.3 Miscellaneous Factors affecting Descriptor Bank Organization

Aitchison quotes several areas in which the practice followed in individual thesauri deviates from the standard to which Descriptor Bank entries must conform. These can be classified into three types of irregularity.

1.3.1 Inter-term Irregularities

Several sources of descriptors do not relate terms solely by means of BT, NT, and RT relationships, e.g.

i) Subject headings, where
   xx means BT or RT (more rarely, NT)
   sa means NT or RT (more rarely, BT)

ii) Classification Systems, where indentation may signify a narrower or related term.

iii) The IBE Thesaurus, where the RT relationship is replaced by reference to a broad facet; not all the terms grouped in such a facet may be formally related.

iv) Certain thesauri which group together terms in topics; not all the terms in a group are formally related to each other by means of the reference symbols.

v) Some thesauri show more than one level of the hierarchy in which a term occurs, e.g. they have BT1 and BT2, or NT1 and NT2 relationships.

vi) Some thesauri, e.g. ISONET, have polyhierarchies, where a term may have more than one BT, and hence occur in more than one hierarchy.

The first four of these irregularities could be dealt with simply by defining an additional relationship, say AT (–?) for Associated Term, solely for use in the Descriptor Bank. This would indicate to the compilers of the Integrated Thesaurus that the term so-marked should be treated with caution in assigning to it a standard thesaural relationship. Alternatively, AT could be subdivided into three relationships, (where P indicates possible) e.g.

BP (<?) – either BT or RT
RP (–?) – possibly RT
NP (>?) – either NT or RT

However if this solution were applied to cases iii and iv, very large numbers of –? terms could be involved, so it may only be suitable for subject headings and classification schemes.

In the case of multi-level hierarchical information, it would be justifiable to ignore the BT2 and NT2 relationships, as these will be indicated under the entries for BT1 and NT1 (as BT1 and NT1). Therefore no information from the thesaurus would actually be lost. The same consideration applies to those thesauri where Top Terms are included for each entry.

Polyhierarchies will exist in the Integrated Thesaurus. However, the choice of the preferred hierarchy in individual thesauri is dependent on the underlying classification scheme. Therefore it may not be particularly useful to distinguish in the Descriptor Bank between preferred and non-preferred hierarchical relationships from individual thesauri, even if this information is represented. The relationships signified in ISONET by asterisks (*>, *–, *<), i.e. cross-references to other parts of the schedule, would be coded in the Descriptor Bank merely as >, – and <.

1.3.2 Intra-term Irregularities

These are a more diverse set of discrepancies between individual thesauri in the representation of single (simple and compound) terms.

i) Character Sets

The candidate thesauri for the Descriptor Bank use a variety of character sets, including upper and lower case, bold type, italics, and special characters, in different combinations. It is essential that a standard be defined for Descriptor Bank entries, since the entries for e.g. FERTILITY, Fertility and fertility must be merged into a single entry, and this can only by achieved if they are identical when compared. The standard must also define a consistent representation for non-English alphabetic characters such as accented vowels in French and Spanish. A common character coding (e.g. ASCII, EBCDIC) will also be required. Conversion between different character sets is a trivial programming task.

ii) Compound Terms

The question of character sets is complicated in the case of compound terms where the relation between the elements of a term is indicated by non-alphabetic characters. Aitchison suggests that e.g. Prisoners, political and Prisoners (Political) should be retained as distinct entries in the Descriptor Bank. It might be more helpful if the two were merged as the difference would appear to be trivial. Whatever decisions are finally made on these matters, they must be formulated and followed precisely, even with regard to spaces before and after non-alphabetic characters, in order to recognise the essential identity of terms.

iii) Singulars and Plurals

These should be left as distinct entries, as they often refer to different meanings of a term.

iv) Spellings

It would not be feasible to attempt to alter the spellings of individual terms (to conform to English rather than American usage) by purely mechanical means. This could only be done by having a complete machine-readable dictionary of differing American and English spellings. However, it might be possible for the software to recognise characteristically American letter sequences (ter, or, ense, ize, ization etc.) and to present terms containing these interactively to an expert for approval, or otherwise, of alteration.

1.3.3 Irregularities in Alphabetical Filing Order

A common alphabetical ordering must be defined on all thesauri for the purposes of merging. Word-by-word filing is preferred to letter-by-letter by Aitchison. The standard must also define an order of filing precedence on all non-alphabetic characters in the Descriptor Bank character set, such as those which indicate non-English characters, and those used in compound terms.

1.4 Requirements of the Descriptor Bank Record Format

Following the above discussion, it is pertinent to enumerate the principal requirements of the logical record that will be used to store information in the Descriptor Bank.

1.4.1 It must be of variable length, capable of accommodating an arbitrarily large number of fields.

1.4.2 The fields themselves should be of variable length, as this will provide for more compact storage and necessitate no truncation of terms, which would be unacceptable.

1.4.3 It must permit the definition of a large number of different field types, in order to accommodate the standard range of thesaural relationships, as many foreign language equivalents as required, and other relations solely for Descriptor Bank use, as described in section 1.3.1.

1.4.4 It must permit the definition of all, or nearly all, field types as repeatable. This applies even to those field types which would not normally be repeatable in a conventional thesaurus, such as class marks and foreign language equivalents, since these may vary from one thesaurus to another.

1.4.5 It must provide a different type of record for each of the outputs defined in 1.4, that is, descriptor, lead-in term, compound descriptor and compound lead-in term.

1.4.6 It must accommodate the storage of the matrix information, either as a bit-string associated with each field of the record, or in some other form which enables the matrix to be easily generated as output.

2. Record Format of Existing Thesauri

The record formats employed in the candidate thesauri for inclusion in the Descriptor Bank were ascertained by providing each thesaurus manager with the computational section of the questionnaire originally developed by Sager et al. (10, 12). The results obtained are presented here in the form of a typology of record formats.

Two principal types of exchange format are used for the thesauri concerned. The types represent alternative

methods of dealing with the essential characteristic of thesaurus structure, i.e. each thesaurus term is associated with an indefinite number of data items (sub-terms etc.)
Type 1. A fixed format record exists for each sub-term, with the main term repeated in as many records as it possesses sub-terms. Examples of thesauri having this format are the National Criminal Justice Thesaurus and the Thesaurus of Psychological Index Terms.
Type 2. A variable format record exists for each term; this consists of three logical segments:

1. Leader – this segment is of fixed length and holds, in a series of fixed length fields, the information relating to the processing of a record in its entirety.
2. Directory – this segment contains the information needed to access the individual fields of segment 3, i.e. the type, length, and offset from the start of the record of each field.
3. Variable fields – this segment contains the data themselves held in variable length alphanumeric fields.

This type of record format is prescribed by ISO (6) and is also used in the ISIS Data Base Management System developed at the International Labour Office (ILO) and maintained at UNESCO (13). Several minor variations in the format exist, e.g. the directory entries may be held as binary numbers (ISIS) or as characters (ISO); and sub-terms of the same type may be held in separate fields, or in a single, 'fragmented', field with sub-field separators. The most significant variation concerns the relationship between a logical record and a physical block; the following sub-types may be defined:

(a) A logical record occupies one physical block. Records may not be longer than a certain maximum length, and shorter records will be padded out to the length of a block. Examples: MACROTHESAURUS, ILO Thesaurus, Population and Family Planning Thesaurus.
(b) A logical record occupies precisely the amount of physical space required, with records placed contiguously, and perhaps across block boundaries. Examples: UNESCO Thesaurus, SPINES Thesaurus.

3. Descriptor Bank Software

This section proposes a record format for the Descriptor Bank, describes the programs required for its creation, and discusses possible software for the development of the Integrated Thesaurus.

3.1 Record Format

Record format type 1, with every sub-term occupying a unique record, is acceptable in terms of storage space only if a given thesaurus term is associated with a small number of sub-terms. Otherwise, the duplication of the main term, with its number etc., in every record corresponding to one of its sub-terms, constitutes an extremely inefficient use of storage space. In the Descriptor Bank, a main term will tend to be associated with a large number of sub-terms, representing the union of the associations of that main term in all constituent thesauri. Therefore a variable record format, with directory, should be adopted for Descriptor Bank use. This format has the added advantage that the directory provides a very compact indication of the number and type of each association in the record, which may be conveniently exploited by display programs.

Contiguous storage of records will be essential in a data base as large as that considered here. It is therefore recommended that the ISIS system and its associated record format (type 2b) be adopted for the Descriptor Bank. This type of record offers other space-saving advantages over the ISO (type 2a) format, such as a shorter leader segment, shorter directory entries, and no field separators in the variable field segment. The ISIS system also provides a data compaction facility, whose use can save approximately 40% of the storage space that would otherwise be required.

Of the requirements for a Descriptor Bank record format listed in 1.4, 1 and 2 (variable record length, variable field length), are satisfield, as described above. A facility for defining a large number of different field types (1.4.3) is provided by the field definition table (FDT), which is used as a parameter to the ISIS software. Repeatability of fields (1.4.4) may be indicated in the FDT. Different types of record (1.4.5) may be specified using the appropriate subfield of the fixed field, and the required output layout for each given in the FDT. Storage of the matrix (1.4.6) is discussed in detail in section 3.2.

By adopting the ISIS system, the task of reformatting individual thesauri is facilitated, since two of the largest, UNESCO and SPINES, already exist in the required format. In addition, the system provides software for the conversion of files from ISO to ISIS format; as the ISO standard is the international one for magnetic tape interchange, this facility should be of considerable value. Other utility programs provided by the ISIS system will be usable for the processing of the Descriptor Bank with little or no modification, though some new programs will be required, as described in the remainder of this section.

3.2 Storage of the Matrix

The matrix is that part of the Descriptor Bank output display that indicates in which of the constituent thesauri the associated item occurs (section 1.2). The information enabling this matrix to be generated must be present in the Descriptor Bank. However, this presents certain problems, as it is not immediately obvious that such information can be easily incorporated within the ISIS record format. Several alternative solutions are proposed below.
3.2.1 The matrix is stored in a separate file from the Descriptor Bank. The latter will be in exactly the same form as a standard ISIS data base file, but a new type of record format must be defined for the matrix file. This possibility can be immediately rejected, as all operations on the Descriptor Bank would have to be performed synchronously with operations on the matrix file, which would require radical reorganization of the data base management software.
3.2.2 The notion of separate files having been rejected, it is apparent that the matrix must be included within the records of the Descriptor Bank. The most compact way of storing a set of binary data items is a bit string. One such bit string, equivalent to one row of the matrix, would be associated with each field of the record. Its length in bytes would be n/8, where n is the number of thesauri included in the Descriptor Bank. A separator character would be necessary to mark the boundary between character and bit information. Thus each field in the record would be lengthened by between 5 and 8

bytes, assuming the eventual inclusion of between 25 and 40 thesauri. Merging thesauri to form the Descriptor Bank would involve the simple operation of setting the bit for the appropriate thesaurus in the bit string when a term match had been found. However, this solution is not really acceptable, as the inclusion of non-alphanumeric information in the variable fields segment of the record could involve extensive modification of system software. It might also affect the use of the data compaction facility, which would more than offset any other saving in storage achieved.

3.2.3 The variable fields segment could be restricted to character information only by storing the matrix implicitly, incorporating a field in the record for a given term for every thesaurus in which that term appears. The thesaurus concerned would be indicated by a single character sub-field at the end of the field, separated form the term by an otherwise unused character, e.g. if 'Person' is used for 'Jail' in two thesauri, A and X, there will be two fields of type USE in the record for 'Jail' (with separate entries in the directory segment) thus:

Prison*A
Prison*X

The non-alphabetic character is provided to maintain the required alphabetical ordering, so that 'Prisons' (plural) occurring as a USE term for 'Jail' in some other thesaurus does not become interposed between the two entries during sorting of fields. An additional field in the record is necessary to hold the main term without a thesaurus code to act as a key for alphabetical sorting of records. Extra processing is necessary to actually generate the matrix at output, by merging the several instances of a term within a given record. More importantly, this solution is very costly in terms of storage. It is obvious that duplication of a term and directory entry in this fashion will fix the size of the Descriptor Bank close to that of the combined sizes of the individual constituent thesauri.

3.2.4 In contrast, if a term is stored once only, with its associated matrix row, the reduction in storage space is proportional to the overlap in content that exists between the constituent thesauri. One way of achieving this was rejected in 3.2.2. However, the matrix row can be represented by a character string rather than a bit string. The most direct representation would provide a byte rather than a bit for each thesaurus, though the increase in field size entailed by this would be unacceptable. An alternative would be for the character string associated with a record field to include only the codes of the thesauri in which the term in that field occurs. A disadvantage of this method is that the matrix information in a record field is of variable length, so that it is no longer accessible by means of arithmetic operations on the corresponding directory entry alone; neither is the term in the field. However, the ISIS system provides a facility for sub-dividing the contents of a single field by means of a sub-field delimiter. This can be used to partition each field into term and matrix sub-fields. It means that a field must be manipulated in order to extract the term alone, for comparison or sorting, and that the character string (matrix row) must be manipulated in order to sequence the thesaurus codes for use in the diplay programs. Both these factors will tend to complicate the merging process.

However, on balance, this representation of the matrix would appear to be the most favourable, though a final judgment on this important design decision should be made only after detailed examination of its affects on all aspects of the host system.

3.3 Conversion Programs

In order to provide the merging program with data in a common format, each thesaurus file must be reformated. Once in ISIS format, all thesauri may be stored by the system. Then, after a simple set of software adjustments, made necessary by the new record types (1.2.3) and field types (1.3.1), any individual thesaurus may be listed using the standard display programs. Access to listings of various thesauri in a common format may be useful for the compilers of the Integrated Thesaurus.

It is quite likely that a separate conversion program will be required for each thesaurus, since each will have idiosyncracies which must be standardized, in content if not in format. This is not a serious problem, as the standardization process will involve only trivial manipulations of the data. For files in ISO format, the conversion program provided by the system can be used as a basis for the individual programs. Similarly, for files with a type 1 format, a general program can be written once, and this will be easily modifiable to deal with the variations between specific thesauri.

In addition to the actual reformatting, the operations which each conversion program must effect are:

1. Standardization of the representation of each term, as described in 1.3.2, including, if necessary, provision of the interactive requests for confirmation of spelling changes.
2. Inclusion of the thesaurus code with all terms, so that each thesaurus has exactly the same record format as the Descriptor Bank. This will facilitate merging.
3. Standardization of the relations between terms, as described in 1.3.1.
4. Exclusion of those sections of the thesaurus which are not to be included in the Descriptor Bank. This may be effected by examination of the classmark or facet indicator associated with each term or group of terms.

In the case of those thesauri which do not provide a classification for terms, exclusion by this method will obviously not be possible, and it will have to be performed manually.

In the case of those thesauri which do not provide a classification code for lead-in terms, the following procedure can be used. All lead-in terms should be converted to standard format and written to one file; descriptors should be checked for relevance, and those to be included converted to standard format and written to another file. The two files can then be organized so that the terms in the USE fields of the records in the lead-in term file can be checked against the main term fields in the descriptor file records. Those lead-in term records where no match is found can be disposed of. Most of the software for doing this will already be available; any extra program which needs to be written will be equally usable for all thesauri undergoing this procedure, as the files concerned will be in the common format.

It is unnecessary that the original thesauri be organized alphabetically. After conversion to ISIS format, the

system software can be used to sort the files into alphabetical order. Small modifications to the sorting programs may be needed to ensure that the terms are filed in the correct order specified by the Descriptor Bank requirements. Of course, if the file is already in this specified order, this sorting stage can be omitted.

Those thesauri which are transferred with a type 1 record format (separate records for each sub-term) should be requested in a form such that all the records for the sub-terms of a given main term are contiguous. Implementations using this format will have standard software to achieve this file organization.

Those thesauri which are only available in machine readable form as files of characters, suitable for direct listing, present no problem with regard to conversion. The use of new line, tab and special characters generally enables thesaural relationships to be extracted unambiguously from such files.

The use of an Optical Character Reader for those thesauri which are currently not implemented on computer should be considered. If this is feasible, then the resulting files can be used as input to conversion programs in the same way as any other.

3.4 Merging Program

Once all thesauri are converted into a common format, and have been alphabetically sorted, the task of merging them into a single file, the Descriptor Bank, is straightforward. The Descriptor Bank is initially constituted using one thesaurus, following which it is merged with each of the others in turn. The basic merging procedure is according to standard principles. The first record of both input files is read and the two main terms are examined. The record whose main term occurs earlier in alphabetical sequence is written to the output file. The next record from the input file which provided the output record is read, and the procedure repeated until both files are empty. If the main term is found in the same field whatever the type of record, and the thesaurus code has been included during conversion, then standard software may be usable for this much of the procedure.

The merge will differ form a conventional merge in that when two records with the same main term are found, special action must be taken. The record type (a sub-field of the fixed field) of the two records must be compared for identity. If this is the case, and the record type indicates a simple term, the information in the two records must be processed in order to update the Descriptor Bank entry with the thesaurus entry. If the type of the record is compound (combined descriptor or factored lead-in term), then each term in the compound must be compared in order to establish whether update should occur. This part of the merging program will need to be specially written.

The thesaurus files should be merged in increasing order of size, as this will keep the size of the Descriptor Bank to a minimum at each stage, and thus reduce the processing required for the entire operation.

3.5 Display, Partitioning and Integrated Thesaurus Development Software

The ISIS system provides a sophisticated facility for specifying the format of hard-copy and VDU output, in the form of a Print Formatting Language. This would permit the printing of the sub-fields of a Descriptor Bank record containing the terms in any required format, with insertion of the appropriate reference symbols into the output as constants. In addition, the table look-up facility provided for the replacement of coded data with the required output might be employed to print the matrix from the ordered string of character codes. It is possible, however, that the table would be of unmanageable size, as it would need to specify all combinations of thesaurus codes. In this case, a procedure for printing the matrix would need to be written.

It is very probable that a complete listing of the Descriptor Bank will be too large for convenient use as a reference tool. Thus it will be necessary to partition the data base into smaller files. The criteria for selection of sub-sets of the Descriptor Bank are not obvious as, of course, no consistent classification exists for all candidate thesauri. However, the inclusion of class marks in the Descriptor Bank will enable choices to be made about its partitioning into specialized sub-sections. First, the broad categories of the BLISS schedule (9) according to which the Descriptor Bank is to be partitioned (1) must be decided by the experts working on the Integrated Thesaurus. These categories will be expressed in terms of the classification schemes of individual constituent thesauri. Then programs will be written which use this information to assign any Descriptor Bank entry with a class mark to a particular sub-section. In the case of a single entry being assigned to more than on sub-section on the basis of several different class marks, this must be resolved manually, (though perhaps interactively), by reference to the BLISS schedule. This must also be done for those entries with no class mark. The advantage of doing this after creating the Descriptor Bank rather than by partitioning individual thesauri, is that any class mark present in an entry may be used for automated partitioning, even if much of the information in that entry originates from unclassified thesauri. Thus the task of manually assigning BLISS class marks to entries is kept to a minimum. Of course, all entries must eventually be classified according to BLISS, but this should be deferred until after partitioning, when it can be done more quickly and accurately by experts in the particular fields of the sub-sections.

At the partitioning stage, the classification is not finalized. After partitioning and examination by experts, any errors in allocation to sub-sections can be corrected, and the classification within a single sub-section refined. The refinement would be assisted by merging the appropriate section of the BLISS classification with the sub-section of the Descriptor Bank, if this had not already been done. A BLISS class mark can then be assigned to those entries in each sub-section which are not already so-marked. When this has been done, the partitioned Descriptor Bank can be reorganized automatically into a classified form, thus bringing together synonymous entries, which will have the same class mark. The relations in the Descriptor Bank will enable decisions about the preferred terms, hierarchies and polyhierarchies of the Integrated Thesaurus to be made. Programs could be provided to assist in these decisions, e.g. by listing all terms with more than one broad term, or all entries which have been assigned the same class mark as another

entry. Other programs, to assist in the definition of the multilingual information needed in the Integrated Thesaurus, e.g. to list all terms with no French equivalent, or more than one French equivalent, could be provided. All programs of this type would be of a very simple nature.

At some time before the reorganization of the Descriptor Bank into its classified display form, it is envisaged that it will be transferred to an installation supporting the ROOT system (3), which is to be used for the Integrated Thesaurus (1). A decision about the precise time that this should be done, and whether the Descriptor Bank should be available on microprocessor at any stage, is left until the characteristics of the Descriptor Bank and its associated software are more accurately known. Once the Descriptor Bank is available under ROOT, interactive software can be provided to eliminate the need for any manual coding of the Integrated Thesaurus entries. This would enable Descriptor Bank entries relevant to a particular Integrated Thesaurus entry to be chosen, one of the terms involved to be designated as the Thesaurus descriptor, and the information merged and reformated as a Thesaurus entry. All this would be under the interactive direction of an expert. This software would be quite complicated, but it would considerably simplify the tedious aspects of compiling the Integrated Thesaurus.

4. Estimates of Resources

In this section, the estimates made as part of the original feasibility study are summarized, in order to give an idea of the size of the task of creating the Descriptor Bank. Table 1 shows the number of terms in each of the candidate thesauri, where this information was available.

THESAURUS NAME	*NUMBER OF TERMS*
ABI/INFORM USER GUIDE	11,000
AGRICULTURAL ECONOMICS AND RURAL SOCIOLOGY	2,300
CHILD ABUSE AND NEGLECT	1,500
ERIC DESCRIPTORS	8,000
HOUSE OF COMMONS LIBRARY	12,400
IBE EDUCATION	2,600
ILO	3,900
MACROTHESAURUS	4,000
NATIONAL CRIMINAL JUSTICE	4,200
POPULATION/FAMILY PLANNING	3,900
PSYCHOLOGICAL INDEX TERMS	6,700
POPULATION MULTILINGUAL	3,000
POLITICAL SCIENCE II	8,000
SCIMP/SCANP	2,000
SPINES	10,500
UNESCO	8,500
URBAN INFORMATION	2,000
TOTAL	84,500
AVERAGE	5,000

Table 1. Numbers of terms in Descriptor Bank Candidate Thesauri

A statistical examination of these thesauri gave an estimate of 1 1/4 million characters (MBytes) for a thesaurus of the average size of 5,000 terms. The storage required for 20 thesauri (in the first instance) would thus be 25 MBytes. This can be taken as an upper bound on the size of the Descriptor Bank. Storage compaction, exclusion of irrelevant sections, and the overlap that exists between thesauri should reduce this figure appreciably. Thus the Descriptor Bank could be stored with ease on an Exchangeable Disc Pack of the sort found at any medium-sized computer installation capable of supporting the ISIS system.

The processing requirements for the creation of the Descriptor Bank were also estimated using typical performance figures for the equipment available at such an installation. The real time necessary for the accessing of the relevant data on secondary storage media, i.e. tape and disc, during the operations of format conversion, sorting, and merging of files was found to be of the order of a few hours. The actual processor time necessary for the manipulation of data in primary memory would be no than 10 minutes. Thus the creation of the Descriptor Bank would pose no significant strain on the resources of a computer installation of the type considered here.

The total programming effort involved in writing format conversion programs, and modifying this ISIS software for the sorting, merging and printing of the Descriptor Bank was estimated at 30 man-days. Again, this estimate is well within acceptable limits.

References

(1) Aitchison, J.: Integrated Thesaurus of the Social Sciences: Design Study. Prepared for UNESCO Division for the International Development of Social Sciences. Letchworth 1981.

(2) Aitchison, J.: Integration of Thesauri in the Social Sciences. Int. Classif. 8 (1981) No. 2, p. 75–85.

(3) British Standards Institution: Root Thesaurus. Hemel Hempstead, 1981.

(4) Dahlberg, I.: Guidelines for the Establishment of Compatability between Information Languages in the Social Sciences. Prepared for UNESCO Division for the International Development of Social Sciences. Frankfurt 1980.

(5) Dahlberg, I.: Toward Establishment of Compatability between Indexing Languages. Int. Classif. 8 (1981) No. 2, p. 86–91.

(6) International Organization for Standardization (ISO): Documentation: Format for Bibliographic Information Interchange on Magnetic Type. ISO 2709, 1973.

(7) Krommer-Benz, M.: Bibliography of Mono- and Multilingual Vocabularies and Thesauri in the Social Sciences. Prepared for UNESCO Division for the International Development of Social Sciences. Infoterm, Vienna, 1981.

(8) Meyriat, J.: Social Science Documentary Languages: A Comparative Analysis. Presented at the Consultative Meeting on the Establishment of an Integrated Thesaurus of the Social Sciences. Paris. 1980.

(9) Mills, J., Broughton, V.: Bliss Bibliographic Classification, vol. 1. Butterworth, London, 1977.

(10) Sager, J.C., Somers, H.L., McNaught, J.: Guidelines for the Establishment of Comparison and Compatbility Matrices between Thesauri in the Social Sciences. Prepared for UNESCO Division for the International Development of Social Sciences. Manchester, 1981.

(11) Sager, J.C., Somers, H.L., McNaught, J.: Thesaurus Integration in the Social Sciences. Part 1: Comparison of Thesauri. Int. Classif. 8 (1981) No. 2, p. 133–138.

(12) Sager, J.C., Somers, H.L., McNaught, J.: Thesaurus Integration in the Social Sciences. Part II: Stages towards Integration. Int. Classif. 9 (1982) No. 1, p. 19–26.

(13) UNESCO: CDS/ISIS: A General Description. Documentation Systems Division, UNESCO, Paris, (1978).

(14) UNESCO: Consultative Meeting on the Establishment of an Integrated Thesaurus of the Social Sciences: Final Report. Division for the International Development of Social Sciences, UNESCO, Paris, 1980.