J.C. Sager, H.L. Somers, J. McNaught
University of Manchester Institute of Science and Technology, England

Thesaurus Integration in the Social Sciences Part III: Guidelines for the Integration of Thesauri

Sager, J.C., Somers, H.L., McNaught, J.: **Thesaurus integration in the social sciences. Pt. III: Guidelines for the integration of thesauri.**
In: Int. Classif. 9 (1982) No. 2, p. 64–70

In this 3rd and last part of a series of articles (Pt. I in Int. Classif. 8 (1981) No. 3, p. 133–138, Pt. II in 9 (1982) No. 1, p. 19–26) guidelines for the successful integration of thesauri are developed in conformity with international standards, taking into account user requirements, the need for an exchange format of data and a suitable database management system. The references to this series have been added to Pt. I. (Acc. to authors)

8. Observations

These guidelines are based on the authors' observations of existing practice, existing standards and guidelines for the compilation and management of multilingual thesauri, which are themselves the results of compromises in existing practices, other relevant standards, consideration of the practicality of their implementation and the general contribution that work towards an integrated thesaurus could make to harmonising linguistic usage in the social sciences.

8.1 General observations

Considering the complexity and expense of thesaurus compilation and maintenance and the inevitable delays in regular updating, which are all multiplied by the number of languages involved in the production of a multilingual thesaurus, alternatives to single subject thesauri should be explored. One such alternative is the merging of thesauri of related subject areas into integrated thesauri as examined in the present study. There are, however, other possibilities for introducing greater economy and efficiency into thesaurus production which merit investigation.

The most obvious means is to compile integrated thesauri in conformity with existing general thesauri and macrothesauri so as to maintain links with other disciplines. But documentation thesauri can also be produced in close correspondence with linguistic or terminological thesauri which are increasingly being developed for a variety of purposes. In principle, hierarchical and associative relationships ought to coincide between these two types of thesauri, even though documentation thesauri may choose to operate on a restricted set of such relations only. Terminological thesauri also contain certain equivalence relationships, i.e. full synonyms and translation equivalents. Thus, a terminological thesaurus is already a useful tool for accessing multilingual free-text based information and documentation systems and may also provide a first approximation to a multilingual controlled language system.

The economy of language achieved through descriptor selection is the result of processes which are overtly declared in the thesaurus by means of USE references. It should be examined which of these processes can be automated so that the indexer or searcher, by inputting a non-descriptor, is automatically referred to the appropriate descriptor. Such a method would appear to be applicable to generic posting, e.g. BARLEY use CEREALS; to full synonyms, e.g. WIRELESS use RADIO; and to quasi-synonyms, e.g. HARDNESS use SOFTNESS. Generic posting and quasi-synonyms are of value only in very specific fields and are therefore unlikely to occur in great numbers in thesauri covering several fields. By making them system-internal the general documentation language would not be affected and would therefore maintain its general usefulness.

Factoring, i.e. analysis of compounds into separate descriptors, and its reverse, synthesis, i.e. the indication of non-descriptors which together form a descriptor, and the resolution of homographs should not be concealed by automation and must remain overtly declared in the thesaurus. Homograph resolution is also a terminological necessity and is carried out by the same methods in both types of thesaurus, i.e. qualifiers and scope notes. Only analysis and synthesis are documentation-specific operations which do, however, coincide with terminological considerations. The factored compound non-descriptor must be a term or else it would not have a place in a thesaurus, the uniterm non-descriptors to be synthesised into a descriptor must also be terms to be meaningful and even the synthesised descriptor should by preference be a term, and in many instances is a compound term.

The coincidences between the considerations that lead to the compilation of both types of thesauri are therefore so great that a documentation thesaurus can be described as an extension of a terminological thesaurus in the direction of greater concentration. As this concentration is largely subject and system specific it is to be considered whether the separate additional effort of producing a multilingual documentation thesaurus can be justified in the light of its much smaller range of applicability. Alternatively it should be considered whether it is preferable to concentrate efforts on producing comprehensive terminological thesauri which, with small extensions, can serve all purposes of multilingual communication including indexing and retrieval.

8.2 Observations on existing standards and guidelines

Since the proposed integrated thesaurus is intended as an indexing and retrieval tool, equivalent to the sectoral thesauri it integrates, and usable alongside these, a structure and format is required which coincides as far as possible with, and which can with time be accepted by, existing and new sectoral thesauri. The specifications of the integrated thesaurus therefore acquire a model character for multilingual thesauri in the social sciences; they can be determined by purely pragmatic criteria, i.e. following the pattern set by highly influential and widely used thesauri, or, alternatively, be based on the good practice established by existing standards and guidelines. Through use of the Checklist we have established that existing guidelines are largely observed and that many variations between thesauri are only the result of the freedom of choice permitted by the guidelines. (It must be remembered that these standards and guidelines are intended for the compilation of thesauri and not for their merging and therefore can provide alternatives which are however unsuitable for integrating thesauri.) For the integrated thesaurus, therefore, choices have to be made among the alternatives permitted by the standards and guidelines and these can be influenced by pragmatic criteria as long as they are consistent.

Choices have to be made at all four levels of description (the knowledge structure, linguistic, formal and computational features) with varying degrees of freedom and ultimate effect on existing thesauri. Since, however, existing guidelines and standards provide no guidance on the broad structure, on some formal and on all computational specifications, the requirements of these levels have to be taken into account additionally. These requirements may themselves narrow the choices available and the degree of variation permissible in future for sectoral thesauri linked to the integrated thesaurus. A distinction also becomes necessary between data structure compatible with merging into an integrated thesaurus and the sector-specific variations which do not affect the integratability of data.

a) On the level of the subject classification, the hybrid nature of most of the thesauri examined favours their integration and the development of a joint broad structure. Class marks are relevant for an integrated thesaurus only to the extent that they affect the address codes.

b) On the linguistic level of description considerable experience exists in multilingual coordination. There are few, if any, fundamental differences of approach and most divergencies should be reconcilable case by case.

c) Variations at the formal level of representation are partly conditioned by computational facilities and partly by the freedom afforded by existing standards and guidelines. For greater ease of use greater uniformity is advisable. Economic considerations may also lead to greater uniformity in future.

d) The computational level of structuring and representation requires careful study and great benefits can be derived from harmonisation to permit economic exchange, transfer and merging of data. At the same time the increasing availability of cheap stand-alone systems and appropriate database management packages may lead to greater diversification. Agreement on compatibility at all levels of hardware, software and datastructure is therefore very important.

For the purpose of integrating thesauri linguistically and formally, and in the interest of international cooperation it is advisable that the variations permitted by existing standards and guidelines be resolved witb reference to the descriptor bank. The ISONET thesaurus (27) defines methods and techniques adopted of particular interest to this study. These are mentioned under 9. and 10. below.

9. Linguistic features

A broad distinction has to be made between the language internal structure and the requirements occasioned by the multilingual nature of a tbesaurus. It is now widely understood that, while thesauri can be and have been translated, it is preferable to plan thesauri multilingually from the start. Every addition of another language, however, requires the total reassessment of all descriptors and it is therefore advisable to construct a thesaurus from the outset with a view to the various languages likely to be required in future.

9.1 Multilingual aspects

The production of genuinely multilingual thesauri, i.e. the use of multilingual source material, the regular adjustment at all stages of the work to the cultural variations in different countries sharing a common language and to the coordination between languages, has led to patterns of factoring descriptors which can serve as models for the integrated thesaurus. It is equally encouraging that some of these thesauri have served as models for other language versions and have therefore proved to some extent their suitability for other languages and cultures. For those sectors of the social sciences for which multilingual thesauri have been developed and in which they are widely used, the multilingual subject headings and descriptors already represent a first compromise necessary for closer cooperation and are therefore steps towards a unified knowledge structure for the purpose of information processing. Their use in retrieval will have an additional, harmonising effect on usage which, by being unobtrusive, can only be beneficial to regularisation of modes of expression in the well-established topics in the social sciences. As multilingual descriptors for new concepts can only be established once the concepts themselves are solidly rooted in documents, such harmonisation cannot be inimical to the creative use of language.

9.2 Non-descriptors

All documentation languages variously control the vocabulary of natural languages by restricting the meaning of terms, often overtly by means of qualifiers or scope notes, by selecting one from a number of synonyms, and by grouping terms (quasi-synonyms) under a convenient descriptor. Further economy in the compression of the vocabulary is achieved through factoring and grammatical form selection.

Genuinely multilingual thesauri, i.e. those which are used by different language communities for indexing of and retrieval from a multilingual database, will have a larger number of descriptors, but, for greater efficiency, also require the entry of a larger number of non-descriptors. In order to allow a greater amount of factoring without losing accessibility it is suggested that an analysis procedure introduced systematically for non-descriptors in the new ISONET Thesaurus be used. This procedure is also found in other thesauri, though less consistently. By this method reference is made from a non-descriptor to two or more separate descriptors; e.g.

Dangerous materials handling
use DANGEROUS MATERIALS
+ MATERIAL HANDLING

Reference can then also be made from the constituent parts, i.e. natural language terms, to a descriptor; e.g.

Petroleum products
+ Fuel storage
= PETROLEUM PRODUCT STORAGE

This procedure is useful for retaining natural language term units, for dealing with uniterm descriptors in one language which are seen as separate units in another and hence factored, and for greater guidance in indexing and retrieval.

9.3 Compound terms and expressions

All thesauri examined contain a high proportion of compound terms and expressions. Various reasons can be given for this preference, some of which are the result of multilingual considerations.

- The availability of precoordinated descriptors facilitates their use in abstracts or analytical summaries;
- Many terms in the social sciences are so well established that their factoring would hinder comprehension;
- Many terms have a linguistic structure in one language which does not permit factoring. As long as the compound is fully established and accepted in the other language it is preferable not to split the term;
- When postcoordination is possible in one language but would lead to ambiguity or loss of meaning in another, the term should be precoordinated in both.

The special factoring rules evolved by BSI for the ISONET Thesaurus seem to be very useful and might be adopted. The introduction of analysis notes also helps to disambiguate compounds and may prove particularly helpful in overcoming different associations suggested by the structure of a language or the knowledge structure of a different cultural system; e.g. in S Social Scientist is listed in the permuted index as follows:

Social Scientist use RESEARCH WORKER + SOCIAL SCIENCES or more specific descriptors as ECONOMIST, SOCIOLOGIST, etc.

By extension S could contain the synthesis notes:

Comparison	Analysis
+ Analysis	+ Comparison
= COMPARATIVE ANALYSIS	= COMPARATIVE ANALYSIS

In multilingual thesauri as in multilingual dictionaries it must be accepted that there will be no absolute equivalence. There are however several ways of dealing with existing differences which need not disturb the structure of the thesaurus:

a) Terms which might be post-coordinated in one language are pre-coordinated because of the other; e.g.

MENSA
UNIVERSITY REFECTORY
RESTAURANT UNIVERSITAIRE
éthnie use GROUP ETHNIQUE

b) The greater use of analysis and synthesis notes, exemplified above

c) The use of scope notes of particular relevance to one language; e.g.

PROFESSEUR SN use in connection with higher education only
COLLEGE SN use SECONDARY SCHOOL if applicable, otherwise use UNIVERSITY COLLEGE

d) When two concepts in one language cannot be distinguished in another, they can be treated as quasi-synonyms; e.g.

Ascension sociale use MOBILITE ASCENDANTE
Décès use MORT
Avenir use FUTUR
Chanson use CHANT

9.4 Word categories and forms

The general recommendation that descriptors should appear in the noun form is widely observed and particularly important for multilingual thesauri; it is more difficult to match adjectives than nouns, as demonstrated in the example of MARRIED.

The use of singular or plural forms is variously prescribed by existing national standards, some of which conflict with the international guidelines. Whereas English language usage is for plurals (BS5723), DIN 1463 requires the use of the singular even for quantitative terms (17). The UNISIST recommendation regarding the plural is observed with one minor exception in M, but not in E and P and S. This deviation from existing rules is explained by the greater ease of abstractors in incorporating singular descriptors into analytical sunmaries which then become searchable and compatible with the theasaurus. It is also adduced that increasingly automated information services accept descriptors both in the singular and the plural and use a truncation procedure for retrieval. It is to be examined whether this greater flexibility, which would avoid the conflict between national and international guidelines can be generally adopted.

Special difficulties arise when singular and plural descriptors coexist; e.g.

DEATHS use MORTALITY P
DEATH E M P S U

9.5 Notes and annotations

The five multilingual thesauri examined consistently use only two types of annotations, scope notes and instructional notes in conjunction with USE references. This approach overloads the functions of scope notes which thereby become definitory, qualificatory and instructional. Since different disciplines and different cultural systems may require parallel systems of relationships, facets may prove useful to accommodate polyhierarchical structures. Only the monolingual U uses a facetted structure. A distinction is also desirable between instructional notes and scope notes; e.g. C uses SN for a dual purpose

RACE RELATIONS SN: Use in connection with race discrimination . . .
COMMUTER SN: Person who has to travel daily . . .

Separate instructional notes can be appended to descriptors and non-descriptors; there are several different types:

a) the analysis of compound non-descriptors into descriptors, as exemplified above,

b) the synthesis on separate non-descriptors into compound descriptors, as exemplified above,

c) reference to a more specific term, e.g.

USE a more specific descriptor
USE in connection with . . . or other descriptors relating to . . .

d) reference to a generic term.

10. Formal features

The formal characteristics of thesauri in machine readable form can only be fully assessed after study of the detailed computational specifications. As the printed versions of thesauri represent only one possible format, which, in addition, may be conditioned by available output devices regardless of its computational structure, it is impossible to describe and assess all formal features on the basis of the printed versions alone. Since the time allocated for this study allowed the authors only to analyse the printed versions, this analysis and the resultant guidelines are by necessity incomplete. Nevertheless, a number of conclusions can be reached on the requirements for a successful integration of thesauri on this level, and these are presented here.

The existing standards and guidelines are only concerned with the printed versions of thesauri, and permit considerable variation which must be restricted for an integrated thesaurus. Such a restriction may also be advisable if in future, as economic considerations would suggest, different sectoral thesauri are produced from a single integrated database.

10.1 Address code notation

The principal function of an address code is to provide a link between the various parts of the thesaurus. It can therefore consist of running numbers only; in a fully structured thesaurus the address code can be indicative of the hierarchical structure and this is the general case in the thesauri examined. The notation can be numeric, as in CEMPS, alphabetic as in the ISONET Thesaurus, or a mixture of both as in U, with the obvious advantages of each such system. If the notation is to be hierarchically expressive it can only be fixed once the classified display is established in its main outline.

The social science specific thesauri examined seen to derive their address codes not from the hierarchical structure of the descriptors, but from some preconceived classification system. The result is a hybrid structure in which the two systems of ordering, i.e. the classification and the interrelationships of concepts, do not match up; e.g. in S 14. (Social Structure), 14200. (Social Stratification)

```
14220.            14230.
(Caste. Slavery)  (Social Class)
CASTE             BOURGEOISIE
ESTATE            LOWER CLASS
NOBILITY          MIDDLE CLASS
SERFDOM           PROLETARIAT
SLAVERY           UPPER CLASS     RT RULING CLASS
                  WORKING CLASS
                  SOCIAL CLASS    NT BOURGEOISIE
                                     LOWER CLASS
                                     MIDDLE CLASS
                                     PROLETARIAT
                                     RULING CLASS
                                     UPPER CLASS
                                     WORKING CLASS
```

```
14250.                    19240.
(Elite. Intellectual)     (Democracy. Dictatorship)
ELITE RT ARISTOCRACY      ABSOLUTISM
ESTABLISHMENT             ANARCHY
 use RULING CLASS         ARISTOCRACY RT ELITE
POWER ELITE               AUTARCHY
 use RULING CLASS         AUTOCRACY
RULING CLASS              .
 BT SOCIAL CLASS          .
 RT POLITICAL ELITE       OLICARCHY
  UPPER CLASS
```

In a fully structured address code, as in U, these terms are much more closely interrelated and their respective relationships made clear by the code and by indentation.

```
R52/64      SOCIAL STRUCTURE
R53/57        SOCIAL STRATIFICATION
R55             CASTE
R55.10            SLAVERY
R56             SOCIAL CLASS
R56.10            WORKING CLASS UF Lower class,
                    Proletariat
R56.20            MIDDLE CLASS UF Bourgeoisie
R56.30            UPPER CLASS
R56.30.10           ARISTOCRACY
R57             ELITE
R57.10            RULING CLASS UF Establishment
                    ARISTOCRACY R56.30.10
R57.10.20           POWER ELITE
R57.30            INTELLIGENTSIA
```

An integrated thesaurus should by preference be hierarchically structured throughout and this structure should be reflected in the address code. In addition every descriptor should have its own address code which indicates the depth of the hierarchy.

10.2 Symbols

It is surprising that multilingual thesauri use symbols based on the English language when there are language independent symbols available and suggested in the existing standards and guidelines. In the past adherence to English was justified by the absence of some symbols on keyboards, but this is no longer the case with a wider international character set. As neutral symbols are becoming more widely available they should be used in preference to English language derived ones. A set of these symbols is contained in BS 5723, another one in ISO (23).

10.3 Displays

The purpose of thesauri being the same, there seems to be little justification for the divergence in display formats. Graphic displays are not as yet widely used and will always be heavily subject dependent, though the proposed experiment of providing a graphic display for the eudised thesaurus may reveal interesting possibilities for the social sciences. Experience has shown that classified and alphabetical displays are both needed with the fullest information possible. Fully structured address codes for individual descriptors and indentation are economical ways of presenting BT/NT relationships in the classified display. A hierarchical display is only useful if reference can be made to it via declared top terms in the other parts. The permuted form is undoubtedly the preferred form for the alphabetical index. The order of elements in the displays is almost entirely uniform.

Whilst there is considerable agreement on the most common types of display and their detailed structure for monolingual thesauri, no such agreement exists on multilingual thesauri and many formats coexist. Bilingual thesauri can generally accommodate two languages in every entry of the systematic part without making the thesaurus unwieldy, as in the case of S. Whether such a method is acceptable and useful for both languages involved is to be examined, as is the question of which part of a thesaurus is regularly used for indexing and which part for retrieval. It has been claimed that indexers prefer to use the alphabetic display, though this practice contravenes the existing guidelines. The widely used Macrothesaurus (M) has a highly developed alphabetic part and only a simple list of descriptors ordered by address codes as the systematic part. Multilingual thesauri in other subject fields also have fully developed alphabetic displays or even dispense with a systematic display altogether.

The multilingual dimension of thesauri can be presented in various ways:

1. Classified display
   a) a full multilingual version as in S; (this may prove cumbersome in cases where several languages are represented)
   b) partially multilingual versions with entry terms only in all languages as in C, E, M and P; (this is possible only if the classified display is not fully structured hierarchically. Indented displays with four languages would probably be unreadable.)
   c) separate versions for each language
2. Alphabetic display
   a) fully multilingual with separate parallel columns for each language as in the E.B.C. Thesaurus (19)
   b) partially multilingual as in M, where only the entry terms are listed in all languages; (such an arrangement requires separate language versions)
   c) separate monolingual displays; (this can be a variation of 2a) above, in that each language in turn provides the ordering sequence for parallel lists)
3. Hierarchic display
   a) as 2a) above; (this is only useful if the top terms of the hierarchies are identified as such in the alphabetic display)
   b) monolingual displays; (this can be arranged in parallel in a permuted order of languages as in 2c) above)
4. Alphabetic index
   a) simple multilingual, each language in turn providing the alphabetic sequence
   b) separately monolingual
   c) permuted multilingual
   d) permuted monolingual

While in principle fully multilingual versions of each type of display seen to be desirable, questions of practical use and also cost should be considered in formulating policy. Which of these formats are preferred in practice will depend on such factors as:

- the nature of indexing and retrieval, e.g.
  - indexing across languages
  - indexing in the language of the document
  - search in the language of possible documents
  - search in one language for documents in another
- the language proficiency of indexers and users
- the degree of automation, e.g. the extent to which a switch in indexing and search language can be automated.

In the absence of any detailed information about patterns of usage no guidelines can be provided for the display formats most desirable for an integrated multilingual thesaurus. It is therefore, suggested that a simple enquiry to establish preferred modes of use for multilingual thesauri be conducted.

10.4 Alphabetisation and filing order

Rules for alphabetisation are needed for all languages in a thesaurus so that each language version can be produced in the order expected by the readers of that language.

10.5 Typography

Some thesauri fully use typography for greater diversification and clarity of presentation. There is, however, considerable variation in the use of block capitals and italics as well as in the use of special characters. It is hoped that an integrated thesaurus will fully use typographical means to increase the efficiency of the printed version of the thesaurus.

11. Technical features

It is understood that for practical reasons the integration of thesauri will be carried out largely by automated means, and that for this reason only such thesauri are being considered which are available in machine-readable form. In the absence of appropriate standards and guidelines and equally, in the absence of a model database structure and management system for an integrated multilingual thesaurus, it is impossible to formulate optimal computational specifications for successful integration.

In principle any two fully automated thesauri are integratable and mergable. If the data records of the two are identical, then merging proves no problem. However, when different data records are involved, then this problem becomes non-trivial. Direct merging of data involves considerable programming effort, which must be repeated for each merge with a different thesaurus. A more reasonable solution would be to adopt an agreed exchange format, which has the effect of minimising the amount of conversion software to be written. In fact, only one suite of conversion programs need be written for each thesaurus. While there may be a need felt for harmonisation of data records used by thesauri, there may be resistance to this due to economic, technical or intellectual reasons. The existence of an exchange or merging format will enable thesauri to retain the data records they find most suitable for their own particular use. Discussions on a suitable magnetic tape exchange format for terminological/lexicographical records have been taking place for some time in ISO. Such a new standard may prove useful for thesauri if the compatibility with appropriate standards in documentation insisted upon by the U.S.S.R. and the U.K. is incorporated.

Though the contents of data records from two or more thesauri may be successfully merged via an exchange format, this does not however overcome the problems posed by differences in the knowledge structure, in the linguistic structure and by the establishment of foreign language equivalents. Such differences can at most be brought to light during a merging sequence. A decision is required whether the effort of integrating one thesaurus with others is economically worthwhile, but this can only be made case by case in the light of the value of the data and the cost of conversion.

Nevertheless, some general observations can be made about merging strategies, about the likely future technical requirements of an integrated multilingual thesaurus and the specifications desirable for permitting such a thesaurus to be hospitable to new developments. Whether made available on-line, or simply stored in machine-readable form for processing purposes, the logical data structure of the thesaurus database should be flexible enough to allow easy updating and editing of records, selective field retrieval and combination of elements. The organisation of schemas and sub-schemas should be such as to permit easily manipulation by future applications programs of unforeseen purpose. An interactive system is recommended for the manipulation of the descriptor bank to permit the detailed study of the consequences of any changes introduced during the merging procedures.

The organisation of the database should be studied in depth: recent advances in database storage systems may be of relevance here, e.g. relational store, associative store, etc. The utilities used to process data must have powerful facilities to handle hierarchically organised data and to ensure proper maintenance of such structures, i.e. deletion, insertion, re-ordering, copying, etc. Output may take the following forms:

- output of selected fields of a record; e.g. term + BT + foreign language equivalent
- output of complete record
- output of selected fields of several records; e.g. to follow a path through a hierarchy
- output of several complete records, required for management purposes only.

Selective field retrieval is necessary for varied output formats and for enabling the on-line searcher to deal with only those fields he needs for his immediate work.

To allow ease of modification and implementation on other host systems, the software used to implement an on-line thesaurus should be of modular design, highly portable, and preferably written in a high-level language suitable for the manipulation of natural language elements. Recent work carried out at the Centre for Computational Linguistics, University of Manchester Institute of Science and Technology in the definition of the requirements of a terminological thesaurus may provide useful additional information (33). If interactive tools are introduced to process data, the same tools can be used to provide the basic components of an interactive on-line user system, or as a front-end processor. It has to be envisaged that thesauri will in future be available on-line and may be consulted interactively, even in question and answer systems, in preference to the printed version. Such a development may lead to the 'hidden' thesaurus, which is implemented as a front-end to a documentation indexing/retrieval system. Users would be unaware of the existence of the thesaurus, in that, e.g. non-descriptors entered as search terms would be automatically and invisibly mapped onto descriptors, including foreign language equivalents, which would then be used by the machine in the search proper. Another possible development may involve the declaration of syntactic links among descriptors used for indexing, along the lines developed by PRECIS for English but with the possibility of extension to a multilingual environment. Finally multilingual thesauri can be used for the automatic translation of keywords in and out of context (see the interesting development of the use of the thesaurus of the Federal German Bundesanstalt für Straßenwesen (12)), as well as for the creation of dictionaries for machine translation systems of full text.

11.1 The suitability of MATER

It is proposed that the suitability of ISO DP 6156 (MATER) (26) as a help in integrating thesauri be investigated. Preliminary examination of the data record fields proposed for use in exchanging terminological/lexicographical information indicates that they are suitable for accepting the types of data associated with thesauri. Certain of these record fields are specifically reserved for thesaurus-type elements; others allow thesaurus-type elements, among others, to be specified.

A brief review of the fields proposed so far suggests the usefulness for handling thesaurus information.

Fields 100–170 are reserved for information on the main entry (variant forms, abbreviations, etc.). Field 180, labelled "search word – expression or words which refer to the main entry", has an obvious implication for thesaurus work. Fields 300–310 give information on subject field and subject classification (optionally this may be in addition to information already encoded in the reference data section of the record). Field 320 indicates the function of the main entry, and the symbols L, N and O are reserved for descriptor, non-descriptor and candidate descriptor, respectively.

Fields 400–440 could presumably be used for scope notes, instructional notes, etc. At present they are described as being reserved for definitions, explanations, examples, footnotes and restrictions. In particular, field 434 is described as "footnote – supplementary comments on the main entry (. . .)".

Fields 500–518 are specifically reserved for information on relationships, catering for several types of BT and NT, i.e. BT, BTP, BTG, NT, NTP, NTG, as well as for antonym, associated term and related term.

Fields 570–576 are reserved for information relating to compounding. Field 573 in particular is set aside for UFC (the reciprocal of a USE of semantic factoring). Field 576 is set aside for "terms which are used in combination in place of the main entry".

Fields of the 600 series are set aside for foreign language terms.

Field 700 is reserved for "facet allocation of main entry" and field 710 may also be useful as it may be used "where the source contains additional notation in respect of data in the fixed data field".

Thus this exchange format would aid in the merging of thesauri, in their comparison, etc., in great part. The status of scope notes and instructional notes, however, in relation to this format must be studied. It would seem that they may be accommodated in the 400 series, though no explicit mention of such notes is made. Also, there is apparently no provision for a term number field. This could be catered for by field 3 of the reference data section of MATER (7.2.2. p 10) which is reserved for "originating agency's internal identification number of the interchange unit". However, it may be felt that an explicit field should be reserved for term number in the data record itself. Given the importance of facetting and synthesis in thesauri, more information on these may be wished in this record format.

11.2 Software for thesaurus management

Various types of software exist for handling mono- and multilingual thesauri. The two main types are:
a) software modules or packages for processing monolingual thesauri, and unstructured thesauri. These include

PROTHE	of Institut français du petrole
PROMETHEE	of INRA
SPLEEN	of CNRS/CDSH
SCRIBE	of French Ministry of Agriculture
INIS	of IAAE

b) software packages providing a fairly complete, or complete DBMS for multilingual thesauri. These include

ASTUTE	of CEC
ISIS	of International Labour Office
SABIR	of Institute Gustave Roussy
EXPLOR	of Laboratoire central des Ponts et Chaussees

Software developed by BSI for the ISONET thesaurus and for its own ROOT thesaurus.

Of these last, two (ISIS and SABIR) are an integral part of information retrieval systems. SABIR was developed in six languages for implementation on CANCERNET. EXPLOR was developed to improve the updating and operation of the trilingual IRRD thesaurus. ASTUTE, then, is the main candidate for the title of fully-fledged DBMS for multilingual thesauri, capable of managing, updating, editing and publishing multilingual thesauri satisfactorily. It is not bound to any one thesaurus, and is meant to aid in the development and maintenance of mono- or multilingual thesauri in general.

Of modular design, ASTUTE comprises some 5,000 COBOL instructions, with a supplementary module of some 500 PL/1 instructions for the preparation of a magnetic tape to drive a photocomposition unit. Apart from the photocomposition module, the system is said to be portable. Up to five languages may be processed per multilingual thesaurus. For each language, a structured multilingual thesaurus may be generated. For display purposes, multilingual versions are printed in combinations of three. Several different output/display formats are possible.

Preliminary perusal of the system specifications and descriptions has revealed several areas where disadvantages appear for handling integrated multilingual thesauri. These are noted below:

- the maximum term length is restricted to 35 characters
- the maximum permitted number of levels in a hierarchy is seven
- accents are ignored
- the repercussions of an input error on the whole thesaurus may be far-reaching, due to the complex file structures involved
- there is a complex correction procedure
- in listings of trilingual thesauri, no scope notes and no non-descriptors are printed
- due to the complex file structures of the system, it is not easy at first sight to appreciate how these, with their various data record formats could be successfully used for data acquisition or exchange. The main dictionary file has 92 character positions. From this are generated several other files, for different purposes, each containing different sets of data.

12. Recommendations

The present article has revealed a number of areas in which further research or coordination seems desirable for the purpose of improving conditions for the successful integration of thesauri.

1. An ISO standard is desirable for the establishment and development of multilingual thesauri. The latest revision of the UNISIST guidelines to that effect, discussed in Somers (24), should be adopted with small additions.

2. ISO 2788 (25) should be revised in the direction of reference to computational requirements.

3. Efforts should be made to establish compatibility between ISO/DP 6156 (Magnetic Tape Exchange Format for Terminological/Lexicographical Records, MATER) and ISO 2709–1973 (Documentation – Format for Bibliographic Information Interchange on Magnetic Tape) (26, 24) so that the former can be used for the proposed multilingual integrated thesaurus for the social sciences. The suitability of ISO/DP 6156 (26) for the exchange of thesaurus data should be established.

4. Practical conditions for international cooperation in the production of multilingual thesauri should be established and offered as models for future work. The introductions to existing multilingual thesauri give details of past practice; these may be compared to determine minimum safeguards for successful cooperation.

5. The use and functions of the different parts of multilingual thesauri for mono- and multilingual indexing and retrieval is neither clearly known nor established. A short enquiry seems advisable so that the most efficient formats can be designed for their different uses. Agreement should then be sought on the presentation of published thesauri, especially on the types of displays and the content structure appropriate for each type.

6. Given the great number of multilingual thesauri with hybrid structures their relative merits or demerits over fully hierarchically structured thesauri should be explored.

7. The link between terminological databanks which contain hierarchically structured data and thesauri should be fully explored, especially the reference to non-descriptors.

8. Aspects of automation of thesaurus lookup should be considered in the process of integrating thesauri.

9. The production of multilingual thesauri will be greatly assisted by the availability of a common international character-set. Efforts in this direction should be supported.

10. Existing clearinghouses for bibliographic information on thesauri should be invited to comment on the Checklist (Part II) with a view to its adoption.

Address:
Prof. Dr. J.C. Sager
UMIST. The University of Manchester
Institute of Science and Technology
P.O. Box 88, Manchester M60 1QD, U.K.