BS 8723-4:2007

Incorporating
Corrigendum No. 1

BRITISH STANDARD

Structured vocabularies for information retrieval – Guide

Part 4: Interoperability between vocabularies

ICS 01.140.20

Publishing and copyright information

ISBN 978 0 580 63073 6

The following BSI references relate to the work on this standard:
Committee reference IDT/2/2
Draft for comment 07/30094123 DC

Publication history

First published December 2007

Amendments issued since publication

Amd. no.	Date	Text affected
C1	30 May 2008	Corrections to formatting of examples in **6.3.1**, **6.4.2** and **A.4.1**

Contents

List of figures

List of tables

Summary of pages

This document comprises a front cover, an inside front cover, pages i to iv, pages 1 to 55 and a back cover.

Foreword

Publishing information

This part of BS 8723 is published by BSI and came into effect on 31 December 2007. It was prepared by Subcommittee IDT/2/2, *Information description, source identification and indexing*, under the authority of Technical Committee IDT/2, *Information and documentation*. A list of organizations represented on this committee can be obtained on request to its secretary.

Supersession

This part of BS 8723 supersedes BS 6723:1985, which is withdrawn.

Relationship with other publications

BS 8723 comprises five parts:

- Part 1: *Definitions, symbols and abbreviations*;
- Part 2: *Thesauri*;
- Part 3: *Vocabularies other than thesauri*;
- Part 4: *Interoperability between vocabularies*;
- Part 5: *Exchange formats and protocols for interoperability*[1)].

Part 1 covers definitions, symbols, abbreviations and other conventions applying to all the parts.

Part 2 covers thesauri, designed for situations in which human indexers analyse documents and express their subjects using thesaurus terms, before searchers retrieve the documents with the same vocabulary.

Part 3 covers other types of structured vocabulary, especially those used in selecting terms or codes for use in subject metadata.

Part 5 sets out the protocols and formats needed for the exchange of vocabulary data.

Information about this document

BS 8723-4 applies to situations in which more than one language or vocabulary is in use, but access to all resources is needed through the one vocabulary chosen by the user. It has a wider scope than BS 6723, which was concerned only with multilingual thesauri, i.e. thesauri presented in more than one natural language; and which required all the language versions to have equal status. BS 8723-4 covers all of the previous ground and extends the scope to:

- thesauri in different dialects of one language;
- different thesauri in a single language;
- situations where a thesaurus interoperates with one or more different types of structured vocabulary, such as classification schemes;
- situations where not all the interoperating vocabularies have the same status and/or function.

[1)] Part 5 is currently in preparation and is expected to be published as a Draft for Development in the first instance.

Use of this document

As a guide, this part of BS 8723 takes the form of guidance and recommendations. It should not be quoted as if it were a specification and particular care should be taken to ensure that claims of compliance are not misleading.

Any user claiming compliance with this part of BS 8723 is expected to be able to justify any course of action that deviates from its recommendations.

Presentational conventions

The provisions in this standard are presented in roman (i.e. upright) type. Its recommendations are expressed in sentences in which the principal auxiliary verb is "should".

Commentary, explanation and general informative material is presented in smaller italic type, and does not constitute a normative element.

Contractual and legal considerations

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

Compliance with a British Standard cannot confer immunity from legal obligations.

Introduction

A demand for interoperability has arisen from the convergence of two trends. Firstly, today's technology permits worldwide interchange of data on a scale not previously imaginable. Secondly, economic pressures dictate that information resources prepared for one application or context should be easily available for other applications. The following cases provide examples of how interoperability is needed to support information sharing in three different scenarios.

a) In a multinational company, knowledge and information gained at one site needs to be accessible to staff in offices around the globe, using their own languages. Hence the need for a vocabulary that functions equally effectively in all the languages used by the company, for the purposes of indexing and searching. The same applies in almost any large organization in which different variations on the same natural language are typically used in different departments.

b) Information produced in the public sector needs to be easily accessible to a variety of audiences. These could, for example, include economists, clinicians and hydrologists, each such group preferring to gain access using its own specialist terminology. However, these terminologies are not intelligible across the discipline boundaries, and in particular not to the lay user. One solution is to index each specialized collection with the appropriate specialized vocabulary, but when the collections are pooled, to provide access via a high-level vocabulary that all users can understand. At first sight this requires double indexing, but the process can be automated if mappings are established between each of the specialized vocabularies and the high-level vocabulary.

c) A third example concerns large collections of data, indexed in past years and decades with different vocabularies in diverse academic institutions. Technology now permits shared access, but searching by subject is impaired by incompatibilities in the indexing vocabularies. Retrospective conversion of the metadata in indexes is not always feasible, but there might be options for converting the terms or codes occurring in search strategies.

A multilingual thesaurus as described by the superseded BS 6723 is often an appropriate solution for case a). Cases b) and c) are more complex, because they require conversion of terms or notations from one vocabulary to another, where the two vocabularies can differ in the selection of concepts and structure. Such conversion may take place at the point of input or at the point of retrieval. In this standard, the mode of operation of a multilingual thesaurus is treated as one approach to tackling a more general requirement for interoperability between vocabularies for the purposes of information retrieval.

1 Scope

This part of BS 8723 gives recommendations for structured vocabularies, when two or more of them are required to interoperate for the purposes of information retrieval. Since any one concept is typically represented by different terms or notations in different vocabularies, the emphasis is on establishing equivalences or other relationships between the various representations of each concept. The case of a multilingual thesaurus, although in one sense this is a single vocabulary, is included within the scope as a special case of interoperability between vocabularies in different languages. There are no comparable provisions for multilingual classification schemes, since their notation applies equally across all languages and the need for equivalence relationships between terms does not arise.

This part of BS 8723 provides recommendations on the way relationships may be established across vocabularies, between the terms and other elements of those vocabularies, so that information may be retrieved from resources indexed with the different vocabularies. It also gives recommendations for managing and maintaining data about the terms and relationships.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

BS 8723-1:2005, *Structured vocabularies for information retrieval – Guide – Part 1: General*

BS 8723-2:2005, *Structured vocabularies for information retrieval – Guide – Part 2: Thesauri*

BS 8723-3, *Structured vocabularies for information retrieval – Guide – Part 3: Vocabularies other than thesauri*

3 Definitions, symbols, abbreviations and other conventions

3.1 Definitions

For the purposes of this part of BS 8723, the definitions in BS 8723-1 and BS 8723-3 apply. The following additional definitions apply to this part of BS 8723. Where a term used in a definition is defined in this standard or in BS 8723-1 or BS 8723-3, that term is italicized.

3.1.1 coined term

new *term* created to express a *concept* for which no suitable *term* exists in the required language

3.1.2 collection

set of information resources that may be accessed by a *structured vocabulary*, whether the items in it are collected in one place or distributed over a network

3.1.3 differentiated mapping
style of *mapping* which differentiates between relationship types such as *equivalence*, associative or hierarchical

3.1.4 equivalence relationship
relationship between two *terms* that both represent the same *concept*

NOTE When two or more such terms *are in the same monolingual* thesaurus*, one of them is designated a* preferred term *and the other(s)* non-preferred term(s)*; the relationship is known as intra-vocabulary equivalence. When both terms are preferred terms in different* thesauri*, the relationship is known as cross-vocabulary equivalence.*

3.1.5 index term
term present in the subject metadata of a document

3.1.6 information retrieval
all the techniques and processes used to provide for identifying items relevant to an information need, from a collection or network of information resources

NOTE Selection and inclusion of items in the collection are included in this definition; likewise browsing and other forms of information seeking.

3.1.7 interoperability
ability of two or more systems or components to exchange information and to use the information that has been exchanged

NOTE Vocabularies can support interoperability by including relations to other vocabularies, by presenting data in standard formats and by using systems that support common computer protocols.

3.1.8 many-to-one mapping
mapping where two or more *terms*, labels or *notations* in one vocabulary are represented by a single *term*, label or *notation* in another vocabulary

3.1.9 mapping
<process>
process of establishing relationships between the *terms*, *notations* or *concepts* of one vocabulary and those of another

3.1.10 mapping
<product of mapping process>
statement of the relationships between the *terms*, *notations* or *concepts* of one vocabulary and those of another

NOTE A mapping generally has a direction, as discussed in Clause **4***.*

3.1.11 multilingual thesaurus
thesaurus using more than one language, in which each *concept* is represented by a *preferred term* in each of the languages, and there is a single structure of hierarchical and associative relationships between *concepts* which is independent of language

3.1.12 one-to-many mapping
mapping in which a single *term*, *notation* or *concept* in one vocabulary is represented by two or more *terms*, labels or *notations* in another vocabulary

3.1.13 one-to-one mapping
mapping in which a single *term*, *notation* or *concept* in one vocabulary is represented by a single *term*, label or *notation* in another vocabulary

NOTE The representations in the two vocabularies may or may not be identical.

3.1.14 precision
measure of retrieval performance defined by R/T, where R is the number of relevant items retrieved and T is the total number of items retrieved

3.1.15 recall
measure of retrieval performance defined by R/N, where R is the number of relevant items retrieved and N is the total number of relevant items in the collection

3.1.16 search term
term forming part of a search query

3.1.17 source language
natural language which serves as a starting point when seeking a corresponding *term* in another language

3.1.18 source vocabulary
vocabulary which serves as a starting point when seeking a corresponding *term* in another vocabulary

3.1.19 target language
language in which a *term* is sought corresponding to an existing *term* in a *source language*

3.1.20 target vocabulary
vocabulary in which a *term* is sought corresponding to an existing *term* in a *source vocabulary*

3.1.21 undifferentiated mapping
style of *mapping* which treats all *mappings* as denoting *equivalence relationships*, although some of them may point simply to the nearest *term* in the *target vocabulary*

3.2 Symbols, abbreviations and other conventions

For the purposes of this part of BS 8723, the symbols, abbreviations and conventions in BS 8723-1:2005, Clause **4** apply. Table 1 repeats the commonest tags appearing in thesauri, and shows the equivalents in some other languages, relevant to examples used in this standard. One additional tag and two symbols are set out in Table 2. The alpha-2 language codes from BS ISO 639-1 are also used in this standard.

Table 1 **Tags and their equivalents in other languages**

Tags in English		Tags in French		Tags in German		Tags in Spanish	
SN	Scope note	NE	Note explicative	D	Definition	NA	Nota de alcance
USE	Use	EM	Employer	BS	Benutze	USE	Use
UF	Use for	EP	Employé pour	BF	Benutzt für	UP	Use por
BT	Broader term	TG	Terme générique	OB	Oberbegriff	TG	Término genérico
NT	Narrower term	TS	Terme spécifique	UB	Unterbegriff	TE	Término específico
RT	Related term	VA	Voir aussi	VB	Verwandter Begriff	TR	Término relacionado

Table 2 **Additional abbreviations and symbols used in mappings**

Symbol	Meaning
EQ	Equivalence marker. The term that follows the tag is the preferred term in a target vocabulary that is closest in meaning to the preferred term preceding the tag, from the source vocabulary.
|	This symbol, a vertical bar, is shown between two or more preferred terms from a target vocabulary, the scopes of which, in sum, best cover the scope of a broader preferred term in the source vocabulary. Each of the target concepts represents *part of the scope* of the source concept. When converting index terms, all the target vocabulary preferred terms should be applied to a record indexed with the source vocabulary term. When converting search statements, the target vocabulary preferred terms should be combined with Boolean OR.
	EXAMPLE
	Source vocabulary → **Target vocabulary**: livestock → sheep | cattle | pigs | poultry
+	This symbol is shown between two or more preferred terms from a target vocabulary which are used in conjunction to represent a compound concept in the source vocabulary. Each of the target concepts represents *an aspect* of the source concept. When converting index terms, all the target vocabulary preferred terms should be applied to a record indexed with the source vocabulary term. When converting search statements, the target vocabulary preferred terms should be combined with Boolean AND.
	EXAMPLE
	Source vocabulary → **Target vocabulary**: women executives → women + executives

4 Structural models for interoperability across vocabularies

4.1 General

At the working level, interoperability is enabled by establishing inter-term relationships, particularly equivalence, for which guidance and recommendations are given in Clauses **6**, **7** and **8**. This clause deals with the overall models within which equivalence and other relationships should be managed.

Three basic models are described in **4.2**, **4.3** and **4.4**.

4.2 Model 1: Structural unity

In the structural unity model, all of the participating vocabularies share exactly the same structure of hierarchical and associative relationships between concepts. It may be represented or expressed in any number of different natural languages, notations or coding systems. A fully multilingual thesaurus should follow this model. All of the languages should have equal status. A collection that has been indexed using any one of the languages can then be searched equally effectively using any of the other languages. Examples are shown in **6.3** and Clause **8**, together with a detailed discussion.

The same model may be used in situations in which at least one of the natural languages is a dialect or sublanguage of another. For example, American English, British English and Indian English may be treated as different languages. Many of the terms may be common to two or three of the languages, but other terms are different. Similarly, the terminology preferred by scientists could be presented as a different language to that of marketing and sales personnel. If the thesaurus is treated as monolingual, one preferred term is assigned to each concept, and the alternative scientific term or dialect term appears as a non-preferred term. Moving to the multilingual model allows equal status to be given to each dialect or sublanguage.

In a multilingual thesaurus, it is usual to establish a direct bi-directional equivalence relationship between corresponding preferred terms in each of the languages. (See examples in **6.3.1**.) The presence of a notation often simplifies data management (see **9.1**).

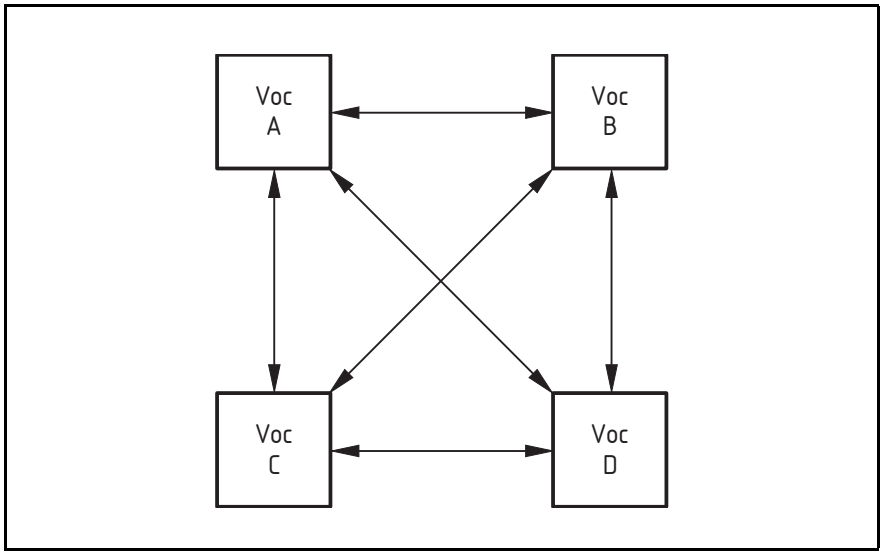
The structural simplicity of this model makes it both feasible and desirable to manage all the terms, notations, captions and relationships between them within one system.

4.3 Model 2: Non-equivalent pairs

The non-equivalent pairs model addresses linkages between two vocabularies that do not share the same structure. As well as differing in scope, language and structure, the vocabularies may include any combination of thesaurus, classification scheme, subject heading scheme, taxonomy, authority list or ontology. Direct mappings should be established between some or all of the terms, notations or categories of one vocabulary and those of the other. The objective of the mappings is to help users to find information in a collection that has been indexed with one of the vocabularies, starting from a search statement that uses terms or notations from a different vocabulary.

This model may be extended to any number of vocabularies/collections, by establishing direct mappings from each vocabulary to each other one. All of the collections can then be searched, using any one of the vocabularies. Figure 1 illustrates the six sets of mappings that are needed to handle four vocabularies, using the model of non-equivalent pairs. The mappings are shown working in both directions, as indicated by the double-headed arrows. However, two-way mappings are sometimes hard to establish. If necessary one-way mappings may be substituted in the same basic model. With one-way mappings, a limitation on applicability of the mappings has to be accepted.

Figure 1 **Model 2 (non-equivalent pairs) as applied to four vocabularies**

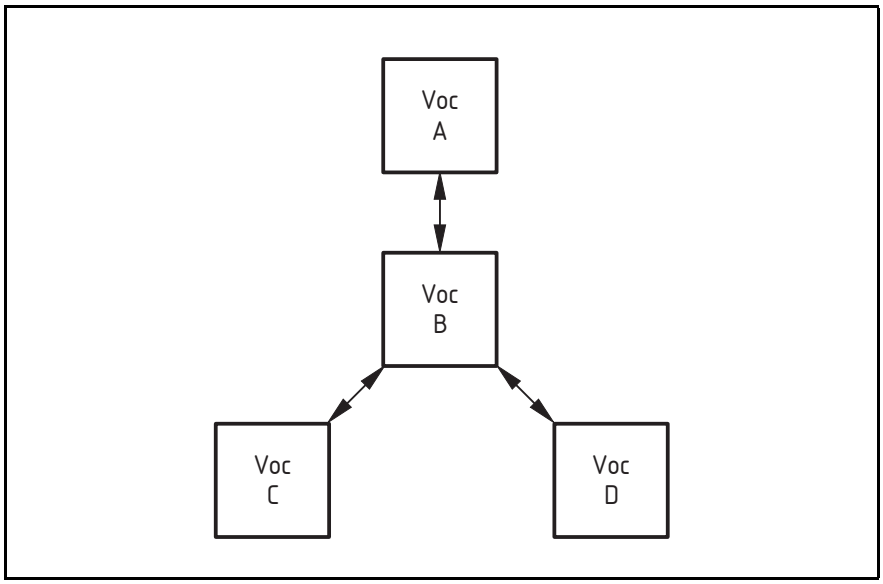


4.4 Model 3: Backbone structure

When more than two vocabularies are required to interoperate, management of the possible combinations becomes complex. It is often convenient to designate one of the vocabularies as the backbone, or basic structure to which each of the subsidiary vocabularies is mapped. Figure 2 shows Model 3, in which vocabulary B serves as a "backbone" to which all the other vocabularies are linked.

Using this model, each concept, category or notation in the backbone vocabulary should be mapped to the corresponding concept(s), category/ies or notation(s) in the other vocabularies. Preferably the mappings work in both directions, as indicated by the double-headed arrows in Figure 2. However, if two-way mappings are not available, one-way mappings may be substituted in the same basic model. (This will restrict its use to mapping in the single available direction.) As Figure 2 shows, only three sets of mappings are needed between four vocabularies using the backbone structure model.

Figure 2 **Model 3 (backbone model) as applied to four vocabularies**



This model enables the backbone vocabulary to be used for searching any of the resources indexed with any of the subsidiary vocabularies. It is usually weaker in supporting use of any one subsidiary vocabulary to search all the indexed resources, since conversion of the search terms is indirect. Whenever a term in one vocabulary is mapped to the nearest term in another, there is a risk of misrepresenting the original concept, and the risk is multiplied by the number of stages through which the conversion needs to pass.

4.5 Application of different scenarios

The three structural models have been enumerated separately but in real applications combinations of them often occur and the boundaries might be blurred.

In choosing between the models, the following factors should be borne in mind.

- The structural unity model (Model 1) is most applicable when a new vocabulary is being prepared for use in more than one language. The absence of an existing indexed collection makes it easier to give each language equal status.
- However, users speaking one language could find it unacceptable to adopt conceptual structures dictated or even influenced by those of another language and in this case Model 2 or Model 3 should be used. The benefit of the latter two models is that there is less artificiality in the use of the user's own language, but the disadvantages are increased complexity for data management, and less interoperability between the indexed collections.
- Models 2 and 3 are appropriate for the reconciliation of vocabularies that have been independently developed and/or have already been applied to collections. Model 2 is applicable particularly when there are only two or three such vocabularies.
- If Model 1 is adopted, work on all the languages should proceed in parallel, with frequent communication between the individuals working in different languages on the same subject areas. In order to achieve equal status between the languages, and to identify the most satisfactory compromise in response to conflicting demands, all members of the editorial team should be flexible in their approach.

Model 1 may also be adopted when an existing thesaurus and/or indexed collection needs access via a language not considered previously. In this case, it is harder to achieve equal status between the languages because translation of terms at this late stage allows less possibility for changing the original thesaurus. This is not the recommended approach, but is sometimes the best option available.

5 Mapping applications in context

5.1 General

In practical applications, interoperability is often achieved by mapping from one vocabulary to another. This clause provides some context as to how these mappings work. Details of the inter-term relationships on which the mappings are based are described in later clauses.

*NOTE Inter-term relationships in multilingual thesauri are discussed in Clause **8** and excluded from this clause.*

The aim of mapping is to switch or augment terms and/or notations, which can be achieved at varying levels of quality. When the mappings in a system are always between exactly equivalent concepts, the effect on retrieval capability is neutral. Assuming that an exact mapping has been effected for every term, there is no effect on either precision or recall, since exactly the same concepts are retrieved, using the alternative vocabulary. However, where the equivalences are inexact or partial, retrieval capabilities are affected. Inevitably some losses arise from the mapping.

Mapping is a two-stage activity. In the preparatory stage, mappings should be established between one set of terms and another. At a later stage, the mappings should be applied to term switching or augmentation. They should be used either to convert index terms that have been applied to a document or to convert terms within a search statement. The intended application should be borne in mind while the mappings are being established.

Once established, mappings may be applied entirely automatically or with human mediation.

Typically, mappings should be established by comparing two vocabularies systematically and for each term in the source vocabulary, finding an equivalent term or combination of terms in the target vocabulary.

Mapping is recommended as likely to give good results when the target vocabulary is rich in the subject area in question, at all levels.

Commonly, the mappings work well in only one direction, and this effect is exacerbated when the vocabularies are of different types. A mapping from vocabulary A to vocabulary B should not be used as a mapping from B to A, without checking its validity in the latter direction.

When bi-directional mappings are required, one approach is to adopt the model of structural unity (see **4.2**) while an alternative is to build a complete set of mappings in each direction.

5.2 The effect of different vocabulary types

5.2.1 General

Mapping is more difficult when the two vocabularies being mapped are of different types (see further discussion in Clause **7**), and in particular when one of the vocabularies contains compound concepts or subject strings as used in pre-coordinate indexing (see **5.6**).

5.2.2 Elements to be mapped

For all types of vocabulary, mappings aim to show relationships between concepts. However, concepts are represented differently in different types of vocabulary.

Table 3 shows the elements which are used to represent the concepts uniquely in different vocabulary types, and which should, therefore, be used in statements of mappings. (See BS 8723-2 and BS 8723-3 for a more complete description of the vocabulary types and elements.)

Table 3 **Elements used to represent concepts**

Vocabulary type	**Concepts represented by:**
Thesaurus	Preferred terms
Classification scheme	Notations
Taxonomy	Category labels, notations or identifiers (see Note)
Subject heading scheme	Terms or pre-coordinated strings (see **5.6**)
Authority list	Terms or identifiers
Ontology	Terms or identifiers

NOTE Different styles of concept representation are used by different taxonomies. Where there is no notation or unique identifier and the category labels are not unique, it is usually necessary to spell out the whole hierarchical path to specify a given concept uniquely. See BS 8723-3.

5.3 Establishing mappings for index terms

The examples that follow in **5.3** to **5.5** have been phrased with reference to terms drawn from a thesaurus. However, they are equally applicable to classification schemes, taxonomies, authority lists and ontologies, provided that notations or identifiers are substituted for terms where necessary. Examples applicable to subject heading schemes can be found in **5.6**.

The context for establishing mappings for index terms is an application in which a collection of documents has been indexed using vocabulary A, but search access is required using terms from vocabulary B. To make this possible, terms from vocabulary B are added to the metadata of each item in the collection, derived by mapping from the vocabulary A terms already present. Terms from either vocabulary can then be used for retrieval. When the mappings are applied, it is important to record which vocabulary each index term comes from, so that there is no confusion if the two vocabularies use the same term for different concepts.

If an exact equivalent is not available, one option is to map to a term that is broader in scope. For example, sheep or cattle may be mapped to livestock. In the context of farming or animal production, this generally gives acceptable results. Items retrieved in a search for livestock are all relevant to that concept. For the users of vocabulary B there is no loss of precision because this vocabulary is not capable of delivering greater precision for that subject.

Mapping from a broader term to a narrower one, for example from livestock to sheep, is usually less satisfactory. A search for sheep could retrieve some items relevant to sheep, together with other items about livestock generally, or about cattle only, or poultry, or pigs. The loss in precision can be considerable.

If the only mapping option available is to a narrower term, all the relevant siblings in vocabulary B should be included. Thus livestock should be mapped to cattle, pigs, poultry and sheep, assuming all these preferred terms are present. It should not also be mapped to chickens or turkeys, if these are present as narrower terms of poultry.

Sometimes the best available option is to map to a term that is neither narrower nor broader, but has an overlapping scope. For example, houseplants might be mapped to potted plants and *vice versa*, even though some houseplants are not in pots, and some potted plants are situated outdoors. A judgement should be made as to whether this will give acceptable retrieval results most of the time, in the context of the intended application.

A more complex type of mapping may be made in the case of multi-word terms (for example, genetically modified wheat) not present in vocabulary B. If vocabulary B contains the separate components of the compound (in this case genetic modification and wheat), these two terms may be used to represent the multi-word term from vocabulary A.

NOTE These examples are summarized in Table 4.

Table 4 **Mapping of index terms**

Original index terms from vocabulary A applied to a document	**Terms from vocabulary B to be applied to the document, derived from a mapping between the vocabularies**	**Notes**
sheep	livestock	Mapping to a broader term, if livestock is the narrowest relevant term in vocabulary B
cattle	livestock	
livestock	cattle pigs poultry sheep	Mapping to several narrower terms where vocabulary B does not contain a term equivalent to livestock
houseplants	potted plants	Mapping to an approximate equivalent
genetically modified wheat	genetic modification wheat	Mapping to two terms to represent a multi-word term

5.4 Mapping of search terms

In establishing mappings of search terms, the intended application is to convert each search term from vocabulary B to the corresponding term or combination of terms in vocabulary A. The same examples as in **5.3** are used to consider the effect, when an exact equivalent is not available.

If a search for sheep is mapped to livestock, precision is lost, but some relevant items might be retrieved among many others. If livestock is mapped to sheep, recall is likely to be poor unless the mapping is to a Boolean combination of all the available types of livestock, for example (sheep OR cattle OR pigs OR poultry). Inclusion of any narrower terms, such as chickens or turkeys, should be an option for the searcher to consider. (Presentation of such options should be generated via relationships within the target vocabulary, not via direct mappings.) The searcher who wants to maximize recall might judge this type of mapping to be very successful; whereas the searcher who wants to restrict results to items dealing with livestock might be disappointed.

In the case of mappings between overlapping terms, such as potted plants and houseplants, there is some loss of both precision and recall. However, acceptable results can be obtained if the overlap is substantial.

Where a multi-word term in a search statement needs to be represented by two or more simpler terms in vocabulary A, this should be done using a Boolean AND. For example, genetically modified wheat may be mapped to (genetic modification AND wheat).

NOTE These examples are summarized in Table 5.

Table 5 **Mapping of search terms**

Original terms from vocabulary A chosen for a search statement	**Terms from vocabulary B to be used in the search, derived from a mapping between the vocabularies**	**Notes**
sheep	livestock	Mapping to a broader term, if livestock is the narrowest relevant term in vocabulary B
cattle	livestock	
livestock	cattle OR pigs OR poultry OR sheep	Mapping to a combination of narrower terms where vocabulary B does not contain a term equivalent to livestock
houseplants	potted plants	Mapping to an approximate equivalent
genetically modified wheat	genetic modification AND wheat	Mapping to a combination of two terms to represent a multi-word term

5.5 Combinations of terms

There is thus a difference between combinations of terms for index term and search term mapping respectively. When mapping index terms, multiple terms are applied to a document without any relationship between them being shown. When mapping terms in a search statement, the terms are combined by appropriate Boolean operators.

In the livestock example in **5.4**, when one term is mapped to a combination of narrower terms, a Boolean OR is applied in the search statement. In the genetically modified wheat example, a Boolean AND is applied. The key difference is that the latter example is of a complex concept represented by a multi-word term, whereas the former example is of a simple term being represented by the sum of members of its class.

To avoid confusion in mapping applications, different symbols, | and +, are used to represent the mappings, as described in Table 2 and further explained in **7.3**.

5.6 Treatment of pre-coordinated strings

*COMMENTARY ON **5.6***
Some subject heading and classification schemes make provisions for pre-coordinated strings (see BS 8723-3:2007, Clause 7) that require extra processing before mappings can be established or applied. Some of these strings may be enumerated in the schedules of the scheme, while others may be created by indexers at the time of indexing a document, by following rules set out in the scheme.

EXAMPLES OF PRE-COORDINATED STRINGS

automobiles -- brakes
automobiles -- clutches
automobiles -- clutches -- handbooks
automobiles -- clutches -- maintenance -- handbooks
automobiles -- handbooks
automobiles -- maintenance
automobiles -- maintenance -- handbooks
clutches -- maintenance
washing machines -- maintenance -- handbooks

Older vocabularies, particularly subject heading schemes, sometimes admit complex concepts in which a degree of pre-coordination is built into the elementary headings. Such headings might include automobile maintenance handbooks *or* space flight on postage stamps*. Pre-processing is not an option in such cases. Such pre-coordination is in practice an obstacle to mapping in either direction, but see advice on treatment of multi-word terms in **5.3** and **5.4**.*

5.6.1 Mapping index terms where the pre-coordinate scheme is the source vocabulary

Three basic options should be considered and a choice should be made in the light of the context of the application, taking into account the arguments that follow.

a) Map the terms or notations representing the separate headings and subdivisions of the pre-coordinate scheme to the target vocabulary.

   In using this sort of mapping, pre-coordinated strings are not mapped, although their component concepts are mapped independently to the target vocabulary.

   EXAMPLE

   The string automobiles -- maintenance is not mapped, although automobiles and maintenance should each be mapped separately.

   This approach achieves a consistent and comprehensive mapping of the constituent elementary concepts, and does not require the rules of the pre-coordinate scheme to be taken into account when applying the mapping. As these rules can be extensive and complicated, they might not be easy to apply in an automated system. This approach has the disadvantage that complex concepts expressed as strings in the source vocabulary are not mapped, even when there is a corresponding complex concept in the target vocabulary, thereby resulting in a loss of precision in retrieval.

b) In addition to a), also map any pre-coordinated strings that are enumerated in the source vocabulary. If possible, map them to pre-coordinated expressions in the target vocabulary, but if not they may be mapped to a combination of terms.

EXAMPLE

The string automobiles -- maintenance should be mapped to the target term automobile maintenance if that exists in the target vocabulary. If it does not exist, the string may be mapped to the combination of terms automobiles + maintenance, as described in Table 2.

This extends the number of complex concepts that are mapped, but might lead to inconsistencies and uncertainty. The scheme might enumerate the string automobiles -- maintenance, but have a rule saying that the subdivision maintenance can be applied to any type of land vehicle. The string bicycles -- maintenance would, however, not be included in the mapping if only enumerated strings were mapped.

The limitations of both a) and b) are difficult to overcome without some human intervention and interpretation at the time of applying the mappings.

c) Sometimes there are good reasons for not using the source vocabulary as the basis of mappings. In order to include additional pre-coordinated strings, mappings may be based directly on the entries in an existing catalogue rather than the nominal source vocabulary.

If a set of catalogue entries is used as the source, an important advantage is that resources do not need to be spent establishing large numbers of mappings that are unlikely ever to be applied. Instead, it becomes feasible to establish mappings from all the pre-coordinated headings that actually occur, and the result is both more comprehensive for the application envisaged, and more accurate at the time of information retrieval. The disadvantage is that the mappings only apply for the mapping of that particular catalogue, and not for other catalogues or indexes or databases that have drawn on the same underlying source vocabulary.

5.6.2 Mapping index terms where the pre-coordinate scheme is the target vocabulary

Where the source vocabulary is a thesaurus (or other vocabulary intended for post-coordinate use) a mapping may be constructed by following the rules of the pre-coordinate scheme. Where possible the source terms should be mapped to individual headings or sub-headings of the target vocabulary, but where this is not possible, a target string may be built to express the source concept.

EXAMPLE

The source term automotive engineering may be mapped to the pre-coordinated string automobiles -- design and construction.

When converting the indexing of documents that have been indexed with a post-coordinate system, pre-coordinated strings cannot usually be derived without human checking or a purpose-built algorithm. For example, if the terms automobiles, upholstery, bodywork, cleaning, costs have been assigned to one document, it is hard to know which combinations to coordinate – automobiles, upholstery and costs? Bodywork, cleaning and costs? Automobiles and cleaning? Rather than attempt ad hoc combinations, mappings should be drawn from those established for the uncombined terms.

5.6.3 Mappings for search statement conversion, where the pre-coordinate scheme is the source vocabulary

Searches for pre-coordinated strings are uncommon if the user has to key them in whole. More commonly, the user selects a pre-coordinated string from a browsable list. This list might be based on the entire source vocabulary, or it might be derived by extracting the terms and pre-coordinated strings in the collection(s) indexed with that vocabulary. To enable the former approach, mappings should be established following the same techniques as in options **5.6.1**a) or **5.6.1**b). For the latter approach, **5.6.1**c) is applicable.

5.6.4 Mappings for search statements, where the pre-coordinate scheme is the target vocabulary

For the conversion of search statements expressed using thesaurus preferred terms, mappings for the individual terms should be provided as described in **5.6.2**. Some of these mappings point to single terms in the pre-coordinate scheme and some to combinations conforming to the rules of the scheme. At the time of applying the mappings, any Boolean operators that are found within the search statements should be retained in the converted statements, but allowance should be made for the occurrence of pre-coordinated headings.

EXAMPLE

A search statement automobile maintenance AND costs may be converted to (automobiles -- maintenance) AND costs.

5.7 Automatic mapping versus human mediation

In general, mapping enables automation or partial automation of the processes associated with information retrieval. Mappings are often employed to extend the retrieval options for a given collection without the cost of additional indexing. Mappings can also be used to speed up human selection where indexing with two vocabularies is needed. When the indexer selects a term from the first vocabulary, the system responds by offering one or more corresponding terms, captions or notations from the second vocabulary. The indexer then confirms appropriate selections and rejects any that are unsuitable.

The advantage of human mediation in mapped indexing is that the indexer can make a judgement about any inexact equivalents, with respect to the document in question. This approach is particularly effective if mappings to broader and narrower terms are offered, as well as nearest equivalents. The disadvantage is cost.

Human mediation is similarly effective in the mapping of search terms. If nearest equivalents are offered, and preferably also broader and narrower terms, the searcher can choose those which best match requirements (not necessarily the same as the preferred terms or notation selected from the source vocabulary).

6 Relationships and mappings across vocabularies and languages

6.1 Types of relationship

6.1.1 General

When relationships are established across vocabularies, they are usually known as mappings. In principle, any type of relationship may be defined and applied as a mapping. In practice, the mappings likely to be useful are determined by the types of vocabulary to be linked. If one of the vocabularies is an ontology, the relationships defined for that ontology are probably needed. If one of the vocabularies is a thesaurus, hierarchical and associative relationships should be considered. The most commonly needed relationship across all vocabulary types is equivalence, as discussed in **6.1.2** and **6.2**.

6.1.2 Intra-vocabulary equivalence versus cross-vocabulary equivalence

BS 8723-2 gives a full description of intra-vocabulary equivalence i.e. the relationship between alternative terms that could be used to designate one concept in a monolingual thesaurus. With intra-vocabulary equivalence, one of the equivalent terms should be selected as a preferred term, while the other terms should have non-preferred term status. In the context of mappings across vocabularies (including the different language versions of one multilingual thesaurus), cross-vocabulary equivalence should be established between preferred terms, identifiers, category labels or notations in different vocabularies that are used to designate the same concept. The cross-vocabulary equivalents should have equal status.

6.2 Degrees of equivalence

*COMMENTARY ON **6.2***
Establishing equivalence is far from straightforward, either across languages or between controlled vocabularies in one language.

The basic need for mapping occurs because of the different languages that different people employ in discourse. These language differences include the obvious one of linguistic variation; and also differences between disciplines (e.g. architecture and civil engineering), and between functions and disciplines (e.g. biochemical research and marketing within a single organization). Likewise, professional and lay people often use different languages, presenting a challenge for navigation and search tools on public-facing websites. In this case it could be necessary to map from a professional language used on an intranet to the equivalent lay terms presented on the Internet.

The obvious translation for a given term from one natural language to another is often valid in most but not all contexts. For example, the term "leg" in English may be translated as "pata" in Spanish, when it refers to the leg of an animal or even a piece of furniture. For the leg of a person, however, the term "pierna" is used instead. Just as two synonyms in one language often have different connotations, it can be difficult to find exact equivalents across languages.

Cultural factors often exacerbate the difficulties. For example a long phrase or sentence might be needed to translate the German term "Berufsverbot", i.e. a law forbidding workers to be employed in certain public sector occupations in the event of their holding radical views, because most other countries do not have the same legal framework.

For non-equivalent vocabularies in one natural language, the first problem is that a different selection of concepts has been made for each. Secondly, even where identical terms are present, they sometimes represent different concepts. For example, the preferred term operations *could have different meanings in a military thesaurus and a medical thesaurus.*

Even when the contexts are similar, there can be subtle differences of scope. For example, the preferred term teachers *in one vocabulary might include teaching staff at universities as well as schools, whereas another vocabulary might have separate preferred terms* lecturers *and* teachers*, with a narrower scope for the latter. The term* public schools *has a very different meaning for American and British contexts, because the educational systems of the two countries are differently organized.*

Both within and between vocabularies, the following degrees of equivalence are often encountered.

a) **Exact equivalence**: In this ideal situation, the target vocabulary contains a concept identical in scope to the concept in the source vocabulary. An equivalence relationship may immediately be established between the corresponding preferred terms, notations or captions.

b) **Inexact equivalence**: In this case, corresponding concepts in the two vocabularies have overlapping scopes. An equivalence relationship may be set up if it is deemed worthwhile, depending on the degree of overlap and the context in which the mappings are to be used.

c) **Partial equivalence**: A concept in one of the vocabularies is broader in scope than a concept in the other. This situation normally calls for the establishment of a hierarchical mapping between the two (see **6.4.2**). However, in contexts where hierarchical mappings are not available, establishing equivalence could be the next-best solution. Alternatively, a combination of two or more preferred terms may be identified in the target vocabulary that together represent the concept in the source vocabulary (see **7.3** and **7.4**).

d) **Non-equivalence**: The target vocabulary does not contain a concept that matches the source vocabulary concept, even partially or inexactly.

The degrees of equivalence in a) to d) should not be treated as distinct relationship types. They are points along the spectrum of possibilities that lie between the extremes of exact equivalence and total absence of equivalence.

The problem is the same, whether it occurs within a multilingual thesaurus or between vocabularies that do not share the same structure. However, as the solutions might be different, these two situations are described separately in Clauses **7** and **8**.

6.3 Equivalence between languages within a multilingual thesaurus

6.3.1 Equivalence between preferred terms

In a multilingual thesaurus, every concept should be represented in every language. Cross-vocabulary equivalences should be established between corresponding preferred terms in the different languages. Once the hierarchical and associative relationships between concepts are established they may be represented as relationships between the corresponding preferred terms in each language.

EXAMPLE

English version	**Spanish version**	**French version**
economic aid	ayuda económica	aide économique
es ayuda económica	**en** economic aid	**en** economic aid
fr aide économique	**fr** aide économique	**es** ayuda económica
UF *economic assistance*		
BT development aid	**TG** ayuda al desarrollo	**TG** aide au développement
NT financial aid	**TE** ayuda financiera	**TS** aide financière
RT economic cooperation financial grants	**TR** cooperación económica donaciones financieras	**TR** coopération économique dons financiers

The example shows a preferred term entry in each language version of a multilingual thesaurus (English, Spanish and French). Hierarchical, associative and intra-vocabulary equivalence relationships in the English version are tagged BT, NT, RT and UF as described in BS 8723-2. Corresponding translated tags are used in the French and Spanish versions. (See Table 1.) Cross-vocabulary equivalents are shown before intra-vocabulary equivalence relationships. They are tagged using the alpha-2 language codes from BS ISO 639-1, which always use lower-case characters. Hierarchical or associative links between terms from different languages are not explicitly shown, because they may be derived via the equivalence relationships.

In a multilingual classification scheme using notation, no mechanism is needed for indicating equivalence, other than the existing tie between a given notation and its captions in each of the languages. If the classification scheme or taxonomy has no notation, equivalence between captions or category labels may be shown just as for thesaurus preferred terms.

6.3.2 Correspondence between non-preferred terms

The number of synonyms or quasi-synonyms for a term usually varies from one language to another. In the example in **6.3.1**, economic assistance is considered to be a synonym for economic aid. However, no synonyms have been found for the Spanish and French preferred terms ayuda económica and aide économique respectively. A multilingual thesaurus does not necessarily require correspondence between non-preferred terms across languages.

It is sometimes possible to establish correspondence between non-preferred terms in the different languages of a multilingual thesaurus. This can be useful when the non-preferred terms are quasi-synonyms conveying a specific aspect of the scope of the preferred term. Non-preferred terms are sometimes established for concepts at the boundaries of the subject scope of the thesaurus because the quantity of information to be described does not warrant a separate preferred term. As and when the subject scope expands, such non-preferred terms may be promoted to preferred terms. In this circumstance it is convenient if the equivalents in all the languages applicable have already been identified.

For example, the term narcissus could be a preferred term in both the English and the French versions of a multilingual thesaurus, with daffodils and jonquille as non-preferred terms for the same concept, in the respective languages. It is possible to establish direct correspondence between daffodils and jonquille, although this is not usually necessary.

6.4 Mappings across structurally different vocabularies

6.4.1 General

In the case of non-equivalent pairs of vocabularies (including each of the pairs in the backbone model), one-to-one mappings between equivalent concepts should be established wherever possible, although these are often difficult to achieve. (For examples of the difficulties, see **6.2**.) Cross-vocabulary equivalence can sometimes be established, but it is not always reliable for the derivation of other types of relationship. In some contexts, it could be necessary to differentiate between mappings to a term equivalent in scope and those to a term which is broader or narrower or just associatively related. In other contexts, it is sufficient to map to an undifferentiated nearest term. A choice should be made between the two styles of mapping (differentiated or undifferentiated respectively) and the corresponding conventions illustrated in **6.4.2** should be followed.

6.4.2 Examples, styles and conventions for non-equivalent pairs of thesauri

The example shows entries from a source vocabulary and two target vocabularies. Table 6 shows how mappings can be established and tagged for the two different styles of mapping. The differentiated style demands much more work, but enables greater flexibility at the application stage.

NOTE The colons shown in Table 6 are optional.

EXAMPLE

Source vocabulary	**Target vocabulary 1**	**Target vocabulary 2**
dairy products	milk products	animal products
UF *milk products*		
BT animal products	**BT** food products	**BT** products
NT butter cheese milk	**NT** ice cream cheeses	**NT** leather meat milk
RT dairies	**RT** dairies	**RT** livestock industry

Table 6 **Two different styles of mapping (mappings from one source vocabulary to two target vocabularies are shown)**

Differentiated mapping		**Undifferentiated mapping**	
dairy products		dairy products	
UF:	*milk products*	UF:	*milk products*
BT:	animal products	BT:	animal products
NT:	butter cheese milk	NT:	butter cheese milk
RT:	dairies	RT:	dairies
VOC1EQ:	milk products	VOC1:	milk products
VOC1BT:	food products	VOC2:	animal products
VOC1NT:	ice cream cheeses		
VOC1RT:	dairies		
VOC2NT:	milk		
VOC2BT:	animal products		

In Table 6, VOC1 and VOC2 are the identifiers of the two vocabularies to which mappings are to be established.

- The left-hand column (differentiated mapping) shows mappings that differentiate between the three standard types of relationship.
- The right-hand column (undifferentiated mapping) assumes that it is sufficient to indicate the one nearest term in each alternative vocabulary.

Hierarchical and associative relationships within the source vocabulary are shown using the standard tags BT, NT and RT as recommended in BS 8723-2.

For mappings, the tag or field label should have two components: firstly the identifier for the target vocabulary, and secondly a tag for relationship type. The identifiers of the source and target vocabularies should be chosen to avoid confusion with any of the standard tags.

Cross-vocabulary equivalence should be tagged EQ, not USE or UF. (The use of USE/UF would cause confusion in this context, because it implies that one of the terms is not a preferred term.)

In the event that the mapping is to an undifferentiated nearest term, the second component of the tag, i.e. "EQ", is unnecessary. Its omission is an acknowledgement that in some cases the nearest mapping is not as close in meaning as might be wished.

If the target vocabulary is multilingual, the identifier for it may additionally include an indicator of the language version within it. In the example, VOC1 should become VOC1en and VOC1EQ should become VOC1enEQ.

The example illustrates the difficult choices required for mapping and also the additional flexibility afforded by the differentiated style.

- Target vocabulary 1 is a reasonably good match for the source vocabulary, with a similar density of preferred terms in the subject area. Mapping is, therefore, fairly straightforward.
- However, as vocabulary 2 is relatively sparse in this area it is hard to decide, for example, whether animal products or milk is closer to dairy products. When the mappings are applied, the results could prove less than satisfactory.
- With the differentiated style, it is unnecessary to decide the nearest match; the user can be offered the choice of a narrower or a broader term, in the light of the particular indexing or search requirement.

The differentiated style has additional advantages if two-way mappings are required.

6.4.3 Mapping non-equivalent pairs of different types

If one or more of the vocabularies is not a thesaurus, the elements to be mapped should be as set out in Table 3. Mappings for relationships other than equivalence are likely to be impractical, and the tags described in **6.4.2** are conventionally used only with thesauri. However, it is often feasible to establish equivalence mappings between pairs of notations from two different classification schemes, or between notations and terms/identifiers/labels, etc. If one of the vocabularies provides for pre-coordinated headings, the options described in **5.6** are available.

7 Establishing equivalence for structurally different vocabularies

7.1 General

In establishing equivalence for structurally different vocabularies, either Model 2 (non-equivalent pairs) or Model 3 (backbone structure) should be chosen (see Clause **4**).

The assumption is made in Clause **7** that the vocabularies are independent and it is not possible to modify them for the purpose of establishing mappings. This limits the range of available solutions. In some cases the vocabularies may also be in different languages, although for simplicity the examples shown here are all in English.

Since Model 1 (structural unity) does not apply it might not be necessary to find a unique mapping for every concept. In other words, two preferred terms in the source vocabulary may sometimes be mapped to the same preferred term in the target vocabulary. Depending on the application, it might not even be necessary to find a mapping for every concept.

Since many applications are for mappings in one direction only, it is sometimes unnecessary to provide for reciprocal mappings in the opposite direction.

NOTE Except where otherwise stated, the examples in this clause assume that the only type of mapping to be established is equivalence.

7.2 Accepting a near-match

An inexactly or partially equivalent concept may be accepted as equivalent if this provides acceptable retrieval results in most cases.

EXAMPLES (near-matches)

Vocabulary 1	**Vocabulary 2**
school buildings	school premises
organizational structure	management structure
devolved government	devolution

7.3 One-to-many cross-vocabulary equivalence

When two vocabularies do not share the same structure, a concept in one of them is not always present in the other. Instead the target vocabulary may have two or more concepts that together make up the sought concept. A one-to-many mapping may then be established. Two different cases arise.

a) A broad but simple concept may be made up of several narrower concepts that are comparable and belong to the same fundamental category.

EXAMPLES

Source vocabulary	**Target vocabulary**
cycles	bicycles | tricycles
hosiery	stockings | socks

b) A complex concept represented by a single term (often a multi-word term as described in BS 8723-2:2005, Clause **7**) in the source vocabulary may be conveyed by the combination of two or more simple terms in the target vocabulary.

EXAMPLES

Source vocabulary	Target vocabulary
executive women	executives + women
rail safety	railways + safety
rail passenger safety	railways + passengers + safety

Although cases a) and b) are for many purposes very similar, some mapping applications need to handle them differently (see **5.3**, **5.4** and **5.5**). For this reason different symbols (| and + respectively) are used to represent the combination.

If the examples in a) and b) are applied in practice, a search for hosiery in a collection indexed with the target vocabulary should be mapped to stockings OR socks. In contrast, a search for executive women should be mapped to executives AND women.

When mapping is applied to the indexing operation, both examples convert in the same way: hosiery should be mapped to stockings and socks; executive women should be mapped to executives and women. When indexing, Boolean operators should not be used; these should apply only in the context of search statements.

It is also important to note that the complex concept mapping examples are safe in one direction only; hence the designation of source and target vocabularies. If a document has been indexed with both executives and women, the inference that it deals with women executives is not reliable; the document might deal with discriminatory treatment of women by male executives. The examples in a) give more reliable results when applied in the opposite direction.

7.4 Many-to-one mappings

When the levels of specificity differ between the vocabularies, several concepts in the source vocabulary may be mapped to the same one in the target vocabulary.

EXAMPLE

Source vocabulary	Target vocabulary
chairs	furniture
desks	furniture
furniture	furniture
tables	furniture

This example could be regarded as the converse of the examples in **7.3**a), except for the assumption that as well as chairs, desks and tables the source vocabulary has a broader term, furniture, corresponding exactly to the nearest equivalent in the target vocabulary.

Many-to-one mappings work in the same way, whether they are applied to index terms or to search terms. However, de-duplication can be useful if the same term is generated twice in one indexing string or search statement.

The many-to-one situation can be more precisely specified if the mappings are differentiated as described in **6.4**. Thus chairs, desks and tables qualify for a hierarchical mapping to furniture, and have no equivalence mapping, as shown here for the same example.

EXAMPLE with differentiated mappings

Source vocabulary	**Differentiated mappings to Target vocabulary**
chairs	TVocBT: furniture
desks	TVocBT: furniture
furniture	TVocEQ: furniture
tables	TVocBT: furniture

The converse of examples as in **7.3**b) is rarely useful. A vocabulary which includes the preferred term rail passenger safety, is likely also to have provided for the concepts of railways, passengers and safety. Therefore an application in which passengers, from the source vocabulary, is substituted by rail passenger safety from the target vocabulary, is hard to imagine.

8 Establishing equivalence between languages in a multilingual thesaurus

8.1 General

The principle of seeking equal status for each language (see **4.2**) should be borne in mind at all stages. See also **12.2**.

In a multilingual thesaurus, every concept should be labelled by a preferred term in each of the languages. (See **4.2**.)

In cases where a suitable term for use as a preferred term is not available in one language, it is often possible to modify the definition of the concept into one that has suitable terms in all the languages. Alternatively an inexact equivalent may be accepted in one language, or a gap in one of the languages may be filled with a loan term or a coined term. Where a concept is represented by a single preferred term in one language, the practice of representing it by a combination of preferred terms in another language should not be considered, as it departs from the model of structural unity (see **4.2**).

A series of practical examples is given in Annex A.

8.2 Accepting a near-match

An inexactly or partially equivalent term may be accepted as a preferred term if this provides acceptable retrieval results in most cases.

For example, incendio in Spanish might be accepted as a label for the concept labelled fires in English, although the former term normally applies to the concept of large-scale fires, arson etc., and not to controlled fires such as those used for heating or cooking. Acceptance is possible only if the terms are likely to be used with the same meaning in all the languages, in the context of the given application.

A scope note should be provided to indicate how the scope differs from what would otherwise be expected. In the case of this example, the scope note of fires might say Use only for large uncontrolled fires.

8.3 Loan terms

8.3.1 It can be hard to find an equivalent in other languages when the concept belongs culturally to users of the source language and is unlikely to arise independently within the community of target language users. The source term may be borrowed to fill the gap in the target language, where it is known as a loan term.

EXAMPLE 1

German	**English**
Schadenfreude	Schadenfreude

EXAMPLE 2

Afrikaans	**English**
veld	veld

8.3.2 A loan term may also be adopted when its translation calls for a long definition or explanation which cannot be used effectively as a preferred term in the target language.

EXAMPLE

English	**German**
teenagers	Teenager **D** Zwischen 13 und 19 Jahren

8.3.3 A loan term is sometimes used in natural language, before a coined translation becomes prevalent. This commonly occurs with scientific discoveries or new technologies. If it appears that the translation is likely to become accepted, this should be adopted as the preferred term, and the loan term should be designated as a non-preferred term.

EXAMPLE

German	**English**
Bremsstrahlung	braking radiation **UF** *Bremsstrahlung* *Bremsstrahlung* **USE** braking radiation

8.3.4 User representatives who are native speakers of the target language should be consulted before any loan term is accepted.

8.4 Coined terms

8.4.1 Terms should be coined only after consultation between indexers, language specialists and subject specialists. Such terms may be created in the following circumstances.

a) The source language term, which represents a new concept to the users of the target language, is for some reason not acceptable as a loan term.

b) The source language term has already been used as a loan term by authors writing in the target language, but the term needs to be replaced, because it is deemed inappropriate or unacceptable. Until the newly-coined term has become established, the loan term should continue to appear in the thesaurus as a non-preferred term.

c) In a thesaurus containing three or more languages, a concept first expressed in one of the languages has already been translated as a coined term in one of the other languages. An indexer working in a third language, faced by a choice between two available loan terms, might prefer instead to coin a term, particularly if that concept is likely to occur in the new target language.

EXAMPLE

German	**English**	**French**
Schlüsselkind	latchkey children	enfant à clé **NE** Enfant dont les parents travaillent pendant la journée et qui est muni d'une clé pour pouvoir rentrer chez lui en sortant de l'école.

8.4.2 Coined terms can be created in the following ways (not in order of preference):

a) literal translation of the source language term or its semantic components;

EXAMPLES

English	**French**
winterization	hiverisation

French	**English**
cuisine minceur	lean cuisine

b) construction of a term or phrase which expresses the general meaning of the source language term;

EXAMPLE

German	**English**	**French**
Bremsstrahlung	braking radiation	rayonnement de freinage

c) the invention of a neologism, which should be as concise as possible to encourage acceptance (these inventions sometimes approximate to literal translations).

EXAMPLES

English	French
steam cracking	vapocraquage **NE** craquage à la vapeur d'eau
turbofans	turbosoufflante
gender mainstreaming	intégration de la dimension de genre

9 Managing mappings and other relationship data

9.1 Management within one system

When vocabularies share the same structure, all the terms, relationships, notations and other elements are usually managed within one system or database. In the case of a multilingual thesaurus, this makes maintaining consistent reciprocal links (see BS 8723-2:2005, Clause **8**) for all relationship types easier, including cross-vocabulary equivalence, and hence encourages consistent indexing and searching. The list of attributes and relationships in BS 8723-2:2005, **12.3.1** usually needs some expansion to hold the multilingual data. For a multilingual classification system, the management of cross-vocabulary links should be achieved by providing each concept or combination of concepts with a unique notation as well as a caption in each language. Multilingual thesauri can optionally also be handled this way.

9.2 Management within two or more separate systems

When managing mappings and other relationship data between non-equivalent pairs maintained in separate databases, two options are possible:

a) the mappings should be maintained in only one of the databases;

b) the mappings should be maintained reciprocally in both databases.

Option b) gives the owners of both databases more flexibility to utilize the mappings as they wish. However, more work is involved. With both options effort is needed to keep the mappings aligned while normal maintenance is in progress.

9.3 Management external to the source and target vocabulary systems

When the mappings data is stored and maintained in a system external to any of the vocabularies, the minimum data to record should be:

a) the set of two or more equivalent preferred terms, notations or captions, one from each of the interoperating vocabularies;

b) the nature of the relationship between them.

In the event that the relationship is not symmetrical (e.g. BT/NT), the direction should also be recorded.

10 Display of mapped vocabularies

10.1 General

No one type of display meets all the likely needs. For some purposes it is sufficient to display only one of the mapped vocabularies, or one of the language versions of a multilingual thesaurus. BS 8723-2:2005, Clause **10** covers provisions for displaying any one language of a multilingual thesaurus.

For some applications the mappings are used mainly by computers. In many applications, humans need not see them at all, if the terms they enter into a search system are converted automatically into the equivalents in another language. In this circumstance, a display of the vocabulary and/or the mappings might be unnecessary.

10.2 Single record display

Even when a continuous display of the vocabulary and mappings is unnecessary, a single record display is often required by the editorial team, particularly while the mappings are being compiled.
BS 8723-2:2005, **10.2** describes the basics for a single record, and should be augmented by mappings using the conventions outlined in **3.2** or **6.4** of the present standard. Figure 3 shows an example. For completeness, many optional features (such as Top Term and Definition) are shown, even though they are likely to be unused in most vocabularies.

NOTE The colons shown in Figures 3, 4 and 5 are optional.

Figure 3 **Single record display in a bilingual thesaurus (English – Spanish)**

pesticides
es: plaguicidas
CC: FF120
SN: Excludes growth regulators and repellents
UF: *fumigants*
TT: agrochemicals
BT: agrochemicals
NT: fungicides
herbicides
insecticides
RT: pests
DEF: A substance for destroying pests
HN: Added 1975-04-01

A similar single-record display should be available for each term in each of the vocabularies. Thus if the editor switches to vocabulary B, the display shown in Figure 4 should be visible.

Figure 4 **Single record display in a bilingual thesaurus (Spanish – English)**

plaguicidas
en: pesticides
CC: FF120
SN: Excluye repelentes y reguladores del crecimiento
UF: *fumigantes*
TT: productos químicos agrícolas
BT: productos químicos agrícolas
NT: fungicidas
herbicidas
insecticidas
RT: plagas
DEF: Agente que combate las plagas del campo
HN: Agregado 1975-04-01

10.3 Multilingual thesaurus displays

10.3.1 Alphabetical displays

The basic style of alphabetical display for a multilingual thesaurus should use the monolingual style described in BS 8723-2:2005, **10.3** and illustrated in BS 8723-2:2005, Figure 5, and be augmented with cross-language mappings as shown in the examples of **6.3.1** in the present standard. Figure 5 shows a brief extract from a bilingual display in this style. A corresponding display should be prepared for each of the languages of the thesaurus.

Figure 5 **Alphabetical display for a bilingual thesaurus**

mosquitoes
  es: mosquitos
  CC: HH120
  BT: pests

pesticides
  es: plaguicidas
  CC: FF120
  SN: Excludes growth regulators and repellents
  UF: *fumigants*
  BT: agrochemicals
  NT: fungicides
      herbicides
      insecticides
  RT: pests

pests
  es: plagas
  CC: HH100
  NT: locusts
      mosquitoes
      slugs
      snails
  RT: pesticides

pigs
  es: cerdos
  CC: HB130
  UF: *hogs*
      porkers
      sows
  BT: livestock

plant products
  es: productos de origen vegetal
  CC: PD100
  NT: cereals
      fruits
      spices
      vegetables
  RT: plants

plants
  es: plantas
  CC: HM100
  RT: plant products

porkers
  USE: pigs

poultry
  es: aves de corral
  CC: HB140
  BT: livestock
  NT: chickens
      ducks
      geese
      turkeys
  RT: eggs

sheep
  es: ovinos
  CC: HB190
  BT: livestock
  RT: wool

Optionally, it might be considered unnecessary to show the language tags for equivalent terms. This has the advantage of saving space in a printed thesaurus, particularly if there are several languages and all the equivalents are printed on one line. However, it is important to avoid confusion by following a consistent sequence for the languages, and differentiating the line of language equivalents typographically (and/or by position) from the other lines. For example, in an English/French/German/Spanish thesaurus some entries might look as shown in Figure 6.

Figure 6 **Alternative layout of entries in an alphabetical display**

```
pigs
  porcin / Schwein / cerdos
  CC:  HB130
  UF:  hogs
       porkers
       sows
  BT:  livestock

plants
  plante / Pflanzen / plantas
  CC:  HM100
  RT:  plant products
```

10.3.2 Systematic displays

For some purposes it is useful to see the structure of the thesaurus laid out in parallel in each of the language versions. Either a hierarchical display as in BS 8723-2:2005, Figure 6, or a classified display as in BS 8723-2:2005, Figure 7, may be chosen. These separate monolingual displays, if they are prepared for each language, could be adequate when laid side by side. Optionally, the language versions may be presented in columns on the same page, so that the correspondence is plainly visible. Figure 7 shows a brief extract from a hierarchical display in this style.

Figure 7 shows notation in a central column for ease of reading across. Notation is optional, but more useful in a multilingual thesaurus than a monolingual one, because it is independent of language. It should be noted, however, that alphabetical ordering of siblings in more than one language simultaneously is usually impossible to achieve, thus giving the impression that one language is dominant. To avoid this problem, some other sequence of siblings may be used throughout, in both languages.

Figure 7 **Hierarchical display for a bilingual thesaurus**

<materials>	FF100	<materiales>
. agrochemicals	FF110	. productos químicos agrícolas
. . fertilizers	FF120	. . fertilizantes
. . pesticides	FF130	. . plaguicidas
. . . fungicides	FF140	. . . fungicidas
. . . herbicides	FF150	. . . herbicidas
. . . insecticides	FM100	. . . insecticidas
. feeds	FP100	. piensos
. fuel		. combustibles
<organisms>	HB100	<organismos>
. livestock	HB110	. ganado
. . cattle	HB120	. . ganado bovino
. . goats	HB130	. . caprinos
. . pigs	HB140	. . cerdos
. . poultry	HB150	. . aves de corral
. . . chickens	HB160	. . . pollos
. . . ducks	HB170	. . . patos
. . . geese	HB180	. . . gansos
. . . turkeys	HB190	. . . pavos
. . sheep	HH100	. . ovinos
. pests	HH110	. plagas
. . locusts	HH120	. . langostas de tierra
. . mosquitoes	HH130	. . mosquitos
. . slugs	HH140	. . babosas
. . snails	HM100	. . caracoles
. plants		. plantas
<products>	PC100	<productos>
. animal products	PC200	. productos animales
. . dairy products	PC210	. . productos lácteos
. . . butter	PC220	. . . mantequilla
. . . cheese	PC230	. . . queso
. . . cream	PC240	. . . crema
. . . milk	PC250	. . . leche
. . eggs	PC260	. . huevos
. . leather	PC270	. . cuero
. . meat	PC300	. . carne
. . wool	PD100	. . lana
. plant products	PD110	. productos de origen vegetal
. . cereals	PD120	. . cereales
. . fruits	PD130	. . frutas
. . spices	PD140	. . especias
. . vegetables		. . hortalizas

10.3.3 Correspondence tables

For some purposes it is sufficient to display only the mappings between the languages in a correspondence table, without any other relationships. Figures 8 and 9 show a convenient layout. A separate table should be prepared for each of the languages, presented in the alphabetical order of the language in question.

NOTE The colons shown in Figure 9 are optional.

Figure 8 **Correspondence table for a bilingual thesaurus (English – Spanish)**

English – Spanish	
agrochemicals	productos químicos agrícolas
agricultural chemicals see: agrochemicals	
animal products	productos animales
butter	mantequilla
cattle	ganado bovino
cereals	cereales
cheese	queso
chickens	pollos
cream	crema
dairy products	productos lácteos
ducks	patos
eggs	huevos
feeds	piensos
fertilizers	fertilizantes
fruits	frutas
fuel	combustibles
fumigants see: pesticides	
fungicides	fungicidas
geese	gansos
goats	caprinos
hens see: chickens	
herbicides	herbicidas
hogs see: pigs	
insecticides	insecticidas
leather	cuero
livestock	ganado
locusts	langostas de tierra
meat	carne
milk	leche
mosquitoes	mosquitos
pesticides	plaguicidas
pests	plagas
pigs	cerdos
plant products	productos de origen vegetal
plants	plantas
porkers see: pigs	
poultry	aves de corral
sheep	ovinos
slugs	babosas
snails	caracoles
sows see: pigs	
spices	especias
turkeys	pavos
vegetables	hortalizas
wool	lana

Figure 9 **Correspondence table for a bilingual thesaurus (Spanish – English)**

Español – Ingles

aves de corral	poultry
babosas	slugs
caprinos	goats
caracoles	snails
carne	meat
cerdos	pigs
cereales	cereals
chanchos véase: cerdos	
combustibles	fuel
crema	cream
cuero	leather
especias	spices
fertilizantes	fertilizers
frutas	fruits
fumigantes véase: plaguicidas	
fungicidas	fungicides
ganado	livestock
ganado bovino	cattle
ganado porcino véase: cerdos	
gansos	geese
herbicidas	herbicides
hortalizas	vegetables
huevos	eggs
insecticidas	insecticides
lana	wool
langostas de tierra	locusts
leche	milk
mantequilla	butter
mosquitos	mosquitoes
ovinos	sheep
patos	ducks
pavos	turkeys
piensos	feeds
plagas	pests
plaguicidas	pesticides
plantas	plants
pollos	chickens
productos animales	animal products
productos de origen vegetal	plant products
productos lácteos	dairy products
productos químicos agrícolas	agrochemicals
puercos véase: cerdos	
queso	cheese

10.3.4 Other displays

Permuted term displays or graphical displays (see BS 8723-2:2005, Clause **10**) may also be prepared, but these should normally be monolingual, with a separate version for each language.

10.4 Displays of mappings between structurally different vocabularies

10.4.1 Alphabetical displays

The same form of display shown in Figure 5 may be used, augmented (as shown in Figure 10) with any additional mappings as outlined in **6.4**.

Figure 10 **Alphabetical display with mappings to two other vocabularies**

agrochemicals
  UF: *agricultural chemicals*
  NT: fertilizers
  pesticides
  VOC1: agrochemical products
  VOC2: agricultural chemicals

animal products
  NT: dairy products
  eggs
  leather
  meat
  wool
  VOC1: animal products
  VOC2: animal products

butter
  BT: dairy products
  VOC1: milk products
  VOC2: animal products

cattle
  BT: livestock
  RT: milk
  VOC1: cattle
  VOC2: cattle

cereals
  BT: plant products
  VOC1: cereals
  VOC2: plant products

cheese
  BT: dairy products
  VOC1: cheeses
  VOC2: animal products

chickens
  UF: *hens*
  BT: poultry
  VOC1: chickens
  VOC2: chickens

cream
  BT: dairy products
  RT: milk
  VOC1: milk products
  VOC2: milk

dairy products
  BT: animal products
  NT: butter
  cheese
  cream
  milk
  VOC1: milk products
  VOC2: animal products

10.4.2 Systematic displays

Systematic displays showing more than one vocabulary are unlikely to be useful when the vocabularies do not share the same structure. However, if the vocabularies have almost identical structures, the display style shown in Figure 7 may be used, leaving gaps where appropriate.

10.4.3 Correspondence tables

Correspondence tables should conform to the display in Figures 8 and 9. There may be gaps in some rows, where mappings do not exist.

10.5 Language and character encoding issues

10.5.1 General

The different languages used in a multilingual vocabulary sometimes use different writing systems. Even languages that share many characters (such as languages using the Latin alphabet) sometimes use different diacritics or other symbols in conjunction with some of those characters. Transliteration schemes are also used to transpose characters in these languages into other writing systems, where there are limitations of printing or display, or in the ability of users to read unfamiliar scripts.

In recent years international standards have been developed to support the computer encoding of a wide variety of languages. These include:

- BS ISO/IEC 8859-1 (Latin-1), which can accommodate most Western European languages; and
- Universal Character Set [1, 2], defined jointly by the UNICODE Standard [3] and BS ISO/IEC 10646, which permits the encoding of most languages of the world.

In particular, the Universal Character Set (UCS) has been specified in other standards such as the HyperText Markup Language (HTML) [4] and the Extensible Markup Language (XML) [5, 6], and has been widely supported by computer manufacturers and software developers. The use of these standards for character encoding promotes interoperability with other systems.

Care should be taken to ensure that all components of a vocabulary management system are compatible and properly configured with the particular languages and character sets. These components include not only the vocabulary management software but also computer hardware for data entry (e.g. keyboards, input software), computer operating systems, and fonts and formats for display purposes (for print publications and computer-based displays). Special consideration should also be given to display issues, filing orders and normalization required for searching and index creation.

10.5.2 Display issues

In each language, characters should appear consistently throughout the vocabulary in a form acceptable to speakers of that language, including diacritics where appropriate. While German speakers sometimes find it acceptable to convert "ä" to "ae" and "ö" to "oe" and so on, in most cases, the form in which characters appear on the printed page or on the computer screen should be the form most familiar to users. Languages that are written from right to left, such as Arabic and Hebrew, usually need special consideration to ensure that text appears correctly displayed and aligned.

10.5.3 Filing orders

Different languages (and even different contexts within the same language) sometimes conventionally require the use of different filing orders, also known as sorting orders. For each language it might be necessary to specify both the characters or diacritics that are ignored for filing purposes and the sequence in which characters should file. For example:

a) the difference between upper and lower case letters is ignored in sorting for English (and in most other languages in the Latin, Greek, Cyrillic and Georgian scripts);

b) in Spanish, the letter ñ files after the letter n, and before the letter o; in French, the tilde (~) is ignored and the letter ñ interfiles with the letter n;

c) in Czech, ch is treated as a single letter coming after the letter h and before i;

d) several alternatives exist for sorting of ideographic characters, as used in China, Japan and the Republic of Korea. In addition, simplified and traditional versions of ideographic characters exist, although the simplified and traditional forms file identically (and are coded identically in BS ISO/IEC 10646);

e) numerical strings are sometimes filed as text, e.g. 1, 10, 102, 11, 120, 2, but other times as numbers, e.g. 1, 2, 10, 11, 102, 120.

For monolingual applications, BS ISO 999 and BS 1749 give basic guidance on filing order suitable for indexes and listings.

Filing orders in multilingual contexts are defined by BS ISO/IEC 14651, and also in the equivalent Unicode Collation Algorithm. However, users may also find it useful to take account of practice in specific languages, which is documented (for languages which use the Latin script) in BS ISO 12199.

Computer software for multilingual vocabularies should be capable of producing conventional filing sequences for each of the different languages in order to facilitate user consultation. If the conventional order cannot be supported, text accompanying each language version of the vocabulary should explain clearly the filing order used.

If displays are required as shown in Figure 7, in which visual correspondence across languages is important, any attempt to use alphabetical order for the sequence of siblings produces unsatisfactory results, for at least one of the languages. Figure 7 illustrates the problem. To avoid giving any language preference, it is necessary either to:

1) produce multiple outputs, one for each language, in which the sequence of all the others is driven by correspondence to the first one; or

2) choose an underlying systematic sequence (e.g. from smallest to largest or from south to north), applying across all the languages.

10.5.4 Normalization for information retrieval

Searching, either within a vocabulary management system or within an information retrieval system using terms taken from a controlled vocabulary, involves matching a string of characters entered by a searcher (search term) with a string of characters extracted from metadata and stored into an index (index term). The search term and the index term sometimes differ in ways that are not considered to be significant, including encoding and the use of upper or lower case characters, diacritics or punctuation. Normalization is the process of converting search and index terms into canonical forms that facilitate string matching and thereby enhance information retrieval by increasing recall.

For example, if a vocabulary management system contains, as a French term, the name of the organization Alliance coopérative internationale, normalization rules applied during the index building process might convert all characters to upper case, and convert accented characters to unaccented equivalents, so that the index contains the string ALLIANCE COOPERATIVE INTERNATIONALE. Similar rules apply when a user enters a search string so that this term record is retrieved by any of the following search strings:

alliance cooperative internationale
Alliance Cooperative Internationale
Alliance Coopérative Internationale
ALLIANCE COOPERATIVE INTERNATIONALE
ALLIANCE COOPÉRATIVE INTERNATIONALE

Regardless of the form the user enters, the two normalized canonical forms match, and the user successfully retrieves the record for the term in the controlled vocabulary.

The above example presents a simple scenario, within a single vocabulary management system. However, normalization rules can also affect the use of spaces (e.g. in place names or pairs of words such as data bases and databases), punctuation characters (particularly hyphens as in words such as fish-breeding), the way that words are extracted, and the use of special symbols such as & (ampersand). (This is described further in BS ISO/IEC 14651, and in the equivalent Unicode Collation Algorithm.)

In rare cases, terms for different concepts might have the same normalized form (e.g. AIDS and aids) and should be treated within the vocabulary as homographs (see BS 8723-2:2005, **6.1.2.2**). Rules for normalization should be made clear to the vocabulary user, and spelling and punctuation rules for the form of terms should take into account the normalization supported by the vocabulary management system.

11 Mapping system functionality

11.1 General

The functions described in BS 8723-2:2005, Clause **11** also apply to multilingual thesauri. The present clause deals only with the additional functions associated with interoperable use of more than one vocabulary or language. Some context is provided in Clause **5**.

11.2 Switching or augmentation of index terms, notations or captions

When preferred terms, notations or captions from more than one vocabulary are added to metadata, the source vocabulary of each such addition should be identified. An appropriate syntax for this purpose is described in standards and guidelines of the Dublin Core Metadata Initiative, for example at http://dublincore.org/documents/dcq-html/ (in which the source vocabulary is known as an encoding scheme).

When the preferred terms to be added are derived from the alternative languages of one multilingual thesaurus, checking by a human operator is unnecessary. However, when terms, notations, etc., are added by using a mapping to a vocabulary with a structure differing from that of the vocabulary of original indexing, human checking should take place.

When checking is carried out, the operator should have reading access to the full content of the document being indexed, as well as to the target vocabulary. It should be possible to browse conveniently from the proposed target term to broader, narrower and related terms in the vocabulary, using the full browse functionality described in BS 8723-2:2005, Clause **11**. It should be easy to deselect the target term and substitute or add others, but not to add terms which are not present in the target vocabulary, unless the application allows uncontrolled terms or candidate thesaurus terms.

11.3 Switching of search terms

For search queries that are presented in natural language, a function which identifies possible preferred terms in the query is very useful. Where the consecutive words in a string might be interpreted as several distinct preferred terms or as one multi-word term preferred term, the options should be presented to the searcher for selection. Where words or word combinations in the query match non-preferred terms, they should be converted to the corresponding preferred terms before seeking confirmation from the searcher.

For search queries expressed using a controlled vocabulary, the search terms should be validated against the vocabulary and any doubts or errors presented to the searcher for resolution. The syntax of statements expressed using Boolean logic should also be checked.

Once the search query has been correctly formulated using a controlled vocabulary, it may be submitted to all the databases or other resources indexed with that vocabulary. If it is to be submitted to resources indexed with another vocabulary, the preferred terms or notations in the query should be converted to the corresponding preferred terms/notations in the other vocabulary, using mappings already established between the two.

Where a one-to-many mapping is indicated (see **7.3**) and Boolean logic is in use, the appropriate operator should be added (see **5.5**).

In some systems the conversion of search statements may be entirely automatic, without the possibility of human intervention. Better retrieval performance can generally be obtained if the searcher is invited to choose between options, and particularly to confirm syntax in the case of one-to-many mappings. When mappings in the differentiated style described in **6.4** are available, there might be several options for each term to be converted. These should be presented to the searcher in a convenient and comprehensible way.

11.4 Expansion of search terms

It is sometimes useful to exploit the relationships in a single thesaurus and/or the mappings across vocabularies, to augment the terms in a query without removing the original ones. This can be particularly appropriate in the context of free-text searching.

If augmentation is completely automated, a measure of the semantic distance applicable to different relationship/mapping types should be developed. Augmentation can then proceed in stages, starting with the nearest terms, until recall has been extended sufficiently.

Better retrieval results are often obtained if the searcher is invited to judge the relevance of additional terms offered via the relationships/mappings.

12 Management of projects for mapping vocabularies and languages

12.1 General

The management of all thesaurus construction projects, monolingual or multilingual, should be conducted in accordance with the provisions of BS 8723-2:2005, Clause **12.**

12.2 Structural considerations

At an early stage in the project, managers should choose between the structural models described in Clause **4**. (See the discussion in **4.5**.) All stakeholders, including users, should be involved in the decision-making process since all the models impose a degree of inconvenience, cost and compromise on them.

12.3 Resources for multilingual projects

The people who work on translation or mapping of vocabularies should ideally have all of the following attributes:

a) a good understanding of each of the natural languages involved;

b) a good knowledge of the subject area of the vocabulary;

c) a good understanding of the difference between normal translation and the identification of equivalents for information retrieval purposes.

Such people are scarce, and are often located in different countries. For this reason it is important to maintain frequent, effective and efficient communication between all team members. Agreement on procedures and formats for unambiguous communication of structured data should be sought.

NOTE Further advice on formats will be given in DD 8723-5, which is currently in preparation.

The project budget should allow for the extra cost of communication overheads, as well as the cost of any specialized software that might be required.

Annex A (informative) Practical examples encountered during preparation of a multilingual thesaurus

A.1 Situation 1

A.1.1 Scenario

No exact equivalent exists in the target language for a concept present in the source language.

EXAMPLE

English	French
teenagers	?

Due to a difference between the English and French names for numbers, teenagers cannot be represented exactly by an existing term in French.

A.1.2 Solution A

Accept the source vocabulary term as a loan term in the target language, and give it a scope note for clarification.

EXAMPLE

English	French
teenagers	teenager **NE** jeune entre 13 et 19 ans

A.1.3 Solution B

Accept a near-match as equivalent.

EXAMPLE

English	French
teenagers	adolescent

A.1.4 Solution C

Reconsider the entries in the source vocabulary and adjust. In this case, the term adolescents might be acceptable as a preferred term instead of teenagers, although it implies a small change of scope. Add non-preferred terms and/or scope notes if necessary.

EXAMPLE

English	French
adolescents **UF** *teenagers*	adolescent

A.1.5 Discussion

There is rarely a unique correct solution. It is important to weigh up the merits of the alternatives and consider what is best for most retrieval situations. Solution C is a clean one if a new thesaurus is being built. However, if the source thesaurus has already been used in indexing a substantial collection, changing a preferred term in it might be unacceptable because it jeopardizes retrieval. Solution B is acceptable if it is judged that the concepts of teenagers and adolescents in the respective languages are very close, although it is advisable to give both terms a scope note to clarify the concept. Solution A might cause confusion if the French preferred term adolescent is to remain alongside teenager, although a case could be made for retaining them as distinct concepts, each with a scope note for clarification. Loan terms are generally to be avoided, but are much more acceptable if they are already coming into common usage.

A.2 Situation 2

A.2.1 Scenario

A concept in the source language is not recognized as a single idea by the users of the target language. Instead, it is regarded as consisting of two or more different concepts, each of which is represented by its own specific term.

EXAMPLE 1 (from a thesaurus on road transport)

English	**German**
skidding	?
?	Rutschen
?	Schleudern

EXAMPLE 2

Spanish	**English**
aislamiento	?
?	insulation
?	isolation

A.2.2 Solution A

Establish equivalence between the single concept in the source language and a coined combination of the more specific terms in the target language. The more specific terms become non-preferred terms.

EXAMPLE 1

English	**German**
skidding	Rutschen und Schleudern **BF** *Rutschen* **BF** *Schleudern* *Rutschen* **BS** Rutschen und Schleudern *Schleudern* **BS** Rutschen und Schleudern

It is important to note that solution A does not offend the principle of structural unity because Rutschen und Schleudern is treated as a single term.

A.2.3 Solution B

Treat the source language term as a homograph, adding qualifying phrases or symbols to distinguish between the more specific meanings recognized in the target language.

EXAMPLE 1

English	German
skidding (forwards)	Rutschen
skidding (sideways)	Schleudern

If this solution is adopted, the general concept represented by the original term in the source language, for example skidding, should be indexed by a combination of the qualified preferred terms, for example skidding (forwards) as well as skidding (sideways). Searchers should use the Boolean expression skidding (forwards) OR skidding (sideways).

A.2.4 Solution C

Avoid using the original term at all, if the source language already has acceptable narrower terms corresponding to what is required in the target language. This is essentially a variation on solution B, which works particularly well for Example 2.

EXAMPLE 2

Spanish	English
aislamiento acústico	acoustic insulation **UF** insulation (acoustic)
aislamiento eléctrico	electrical insulation **UF** insulation (electrical)
aislamiento físico	isolation
aislamiento térmico	thermal insulation **UF** insulation (thermal)

In this particular case, if non-preferred term entries in English are provided for the inverted form of each type of insulation, a preferred term for the broader term insulation is not needed because all the main types of insulation are provided for. In Spanish, terms for the main types of insulation are acceptable as they are in common usage, and the only term that looks slightly artificial is aislamiento físico. A scope note should be provided in case this causes any confusion.

A.2.5 Solution D

As in solution A, coin an artificial term for the combined concept. However, retain the more specific concepts as narrower terms in all the languages.

EXAMPLE 1

English	German
skidding **NT** skidding (forwards) **NT** skidding (sideways)	Rutschen und Schleudern **D** Zu benutzen, wenn ein Dokument sowohl Rutschen als auch Schleudern behandelt **UB** Rutschen **UB** Schleudern
skidding (forwards) **BT** skidding	Rutschen **OB** Rutschen und Schleudern
skidding (sideways) **BT** skidding	Schleudern **OB** Rutschen und Schleudern

A.2.6 Solution E

Adopt the source language term as a loan term in the target language, and the target language terms as loan terms in the source language, adding scope notes, if necessary, to all loan terms.

EXAMPLE 1

English	German
skidding **NT** Rutschen **NT** Schleudern	skidding **D** Lehnwort aus dem Englischen, bedeutet sowohl Rutschen als auch Schleudern **UB** Rutschen **UB** Schleudern
Rutschen **SN** Skidding forwards; loan term adopted from German **BT** skidding	Rutschen **OB** skidding
Schleudern **SN** Skidding sideways; loan term adopted from German **BT** skidding	Schleudern **OB** skidding

A.2.7 Discussion

None of the solutions is completely satisfactory, and the acceptability varies from one example to another. Solution C is very satisfactory for Example 2, but does not work well for Example 1.

Solution A requires German-speaking users to accept an artificial combination term. Furthermore, they are unable to use the simple terms Rutschen or Schleudern, because these are defined as non-preferred terms.

Solution B looks less "foreign" to all users, but English speakers have to think in an unfamiliar way. When indexing a document that deals with skidding, it can be difficult or impossible to decide whether the skid is forwards or sideways or both. The problem can be overcome, however, by applying both preferred terms. At the search stage, there is flexibility to search for either of the more specific terms, or a Boolean OR combination of both.

Solution D, combining the best of solutions A and B, is more thorough than either of these. However, it also inherits the disadvantages of A and B – the need to accept some artificial terms at some level in all the languages.

Solution E avoids the artificial terms, but users have to accept some loan terms in all languages. For the example given, the loan terms look unfamiliar and therefore possibly unacceptable. But in other cases solution E might be more acceptable.

A.3 Situation 3

A.3.1 Scenario

A complex concept represented by a single term in the source language is represented by two or more separate preferred terms in the target language, and the source language concept can equally be represented by separate terms which are exact equivalents to the existing terms in the target language.

EXAMPLE 1

English		**French**
solar heating	{	chauffage énergie solaire

EXAMPLE 2

German		**English**
Schäferhunderziehung	{	sheepdogs training

A.3.2 Solution A

Treat the single term in the source language as a non-preferred term, to be represented by a combination of two preferred terms in the same language. Establish equivalence between these and the corresponding preferred terms in the target language.

EXAMPLE 1

English	**French**
heating	chauffage
solar energy	énergie solaire
solar heating **USE** heating **+** solar energy	

EXAMPLE 2

German	**English**
Schäferhund	sheepdogs
Erziehung	training
Schäferhunderziehung **BS** Schäferhund **+** Erziehung	

A.3.3 Solution B

Where the meaning of the preferred term for the complex concept in the source language is lost or distorted by splitting it, retain this term and represent it with a combination of preferred terms in the target language. Equivalents of the component preferred terms should also be included in the source language.

EXAMPLE

English	**French**
solar heating	chauffage + énergie solaire
heating	chauffage
solar energy	énergie solaire

Solution B breaks the rule of structural unity, because there is no one preferred term equivalent to solar heating. For the human user, this approach appears straightforward. However, the software used to maintain the vocabulary database might have difficulty in representing the concept satisfactorily in all languages. The conventions for representing relationships break down completely if solar heating has narrower terms or related terms. However appealing it might seem, solution B is unacceptable.

A.3.4 Solution C

Coin a term in the target language which serves as an equivalent to the complex term in the source language, adding a scope note to the newly-coined term if its meaning is not obvious.

EXAMPLE 1

English	French
solar heating	chauffage solaire

EXAMPLE 2

German	English
Schäferhunderziehung	sheepdog training

A.3.5 Discussion

Solution B is ruled out because it does not conform to structural unity. Solution C is not recommended unless the newly-coined term is likely to be known and sought by users. Solution A is preferred, unless it causes unacceptable retrieval difficulties for a collection already indexed with the single term.

A.4 Situation 4

A.4.1 Scenario

The source vocabulary hierarchy has in it an intermediate level not present in the target vocabulary:

EXAMPLE 1

French hierarchy	English hierarchy
bétail	livestock
gros bétail	–
boeuf	cattle
cheval	horses

EXAMPLE 2

German hierarchy	English hierarchy	French hierarchy
Gastropode	gastropods	gastéropodes
Schnecke	–	–
Gehäuseschnecke	snails	escargot
Nacktschnecke	slugs	limace

A.4.2 Solution A

Coin a term that captures the same meaning. A scope note should usually be added.

EXAMPLE 1

French	**English**
gros bétail	large animal livestock
TG bétail **TS** boeuf	**SN** Equivalent to the French term 'gros bétail'
TS cheval	**BT** livestock
	NT cattle
	NT horses

EXAMPLE 2

German	**English**	**French**
Schnecke	slugs and snails	escargot et limace
OB Gastropode **UB** Gehäuseschnecke **UB** Nacktschnecke	**SN** Equivalent to the German term 'Schnecke'	**NE** Equivalente à l'allemand 'Schnecke'
	BT gastropods	**TG** gastéropodes
	NT snails	**TS** escargot
	NT slugs	**TS** limace

A.4.3 Solution B

Adopt the intermediate term occurring in the source language as a loan term in the target language, adding a scope note to this term where its meaning is not self-evident.

EXAMPLE 1

French	**English**
bétail	livestock
TS gros bétail	**NT** gros bétail
gros bétail	gros bétail
TG bétail	**SN** Means both cattle and horses; loan term adopted from French
TS boeuf	**BT** livestock
TS cheval	**NT** cattle
	NT horses
boeuf	cattle
TG gros bétail	**BT** gros bétail
cheval	horses
TG gros bétail	**BT** gros bétail

EXAMPLE 2

German	English	French
Schnecke **UB** Gehäuseschnecke **UB** Nacktschnecke	Schnecke **SN** Means both slugs and snails; loan term adopted from German **NT** snails **NT** slugs	Schnecke **NE** Emprunt de l'allemand. Signifie à la fois escargot et limace **TS** escargot **TS** limace
Gehäuseschnecke **OB** Schnecke	snails **BT** schnecke	escargot **TG** schnecke
Nacktschnecke **OB** Schnecke	slugs **BT** schnecke	limace **TG** schnecke

A.4.4 Solution C

Make the intermediate term a non-preferred term in the source language, pointing to whichever preferred term is usually regarded as closest in meaning. This may be either a broader or a narrower term.

EXAMPLE 1: assuming that "gros bétail" is regarded as closer in meaning to "bétail" than to either "boeuf" or "cheval"

French	English
bétail **EP** *gros bétail* **TS** boeuf **TS** cheval	livestock **NT** cattle **NT** horses
gros bétail **EM** bétail	

EXAMPLE 2: assuming that Schnecke is regarded as referring more often to Gehäuseschnecke than to Nacktschnecke or to Gastropode.

German	English	French
Gastropode **UB** Gehäuseschnecke **UB** Nacktschnecke	gastropods **NT** snails **NT** slugs	gastéropodes **TS** escargot **TS** limace
Gehäuseschnecke **BF** *Schnecke* **OB** Gastropode	snails **BT** gastropods	escargot **TG** gastéropodes
Nacktschnecke **OB** Gastropode	slugs **BT** gastropods	limace **TG** gastéropodes
Schnecke **BS** Gehäuseschnecke		

A.4.5 Discussion

Solutions A and B depend on judgements as to the acceptability and comprehensibility of the coined and loan terms respectively, which vary from case to case. The context can also influence the choice; for Example 2 an acceptable solution for a zoological or horticultural context might become unacceptable if the context includes culinary subjects. Solution C avoids any inconvenience in the target language, but it involves some loss of retrieval capability in the source language.

Bibliography

Standards publications

For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

BS 1749, *Recommendations for alphabetical arrangement and the filing order of numbers and symbols*

DD 8723-5, *Structured vocabularies for information retrieval – Guide – Part 5: Exchange formats and protocols for interoperability*

BS ISO 639-1, *Codes for the representation of names of languages – Part 1: Alpha-2 code*

BS ISO 999, *Information and documentation – Guidelines for the content, organization and presentation of indexes*

BS ISO 12199, *Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet*

BS ISO/IEC 8859-1, *Information technology – 8-bit single-byte coded graphic character sets – Part 1: Latin alphabet No. 1*

BS ISO/IEC 10646, *Information technology – Universal Multiple-Octet Coded Character Set (UCS)*

BS ISO/IEC 14651, *Information technology – International string ordering and comparison – Method for comparing character strings and description of the common template tailorable ordering*

Other documents

[1] WORLD WIDE WEB CONSORTIUM. *Character model for the World Wide Web 1.0: Fundamentals*. W3C recommendation, 15 February 2005.[2)]

[2] WORLD WIDE WEB CONSORTIUM. *Character model for the World Wide Web 1.0: Normalization*. W3C working draft, 27 October 2005.[3)]

[3] UNICODE CONSORTIUM, ed. JOAN ALIPRAND et al. *The Unicode standard, version 4.0*. Boston, MA: Addison-Wesley, 2003. ISBN 0-321-18578-1.[4)]

[4] WORLD WIDE WEB CONSORTIUM. *HTML 4.01 specification*. W3C specification, 24 December 1999.[5)]

2) Latest version available at http://www.w3.org/TR/charmod. Accessed 20 September 2007.

3) Latest version available at http://www.w3.org/TR/charmod-norm. Accessed 20 September 2007.

4) Online version available at http://www.unicode.org/versions/Unicode4.0.0/. Accessed 20 September 2007.

5) Latest version available at http://www.w3.org/TR/html401. Accessed 20 September 2007.

[5] WORLD WIDE WEB CONSORTIUM. *Extensible Markup Language (XML) 1.0 (fourth edition)*. W3C recommendation, 16 August 2006, edited in place 29 September 2006.[6)]

[6] WORLD WIDE WEB CONSORTIUM. *Extensible Markup Language (XML) 1.1*. W3C recommendation16 August 2006, edited in place 29 September 2006.[7)]

6) Latest version available at http://www.w3.org/TR/REC-xml. Accessed 20 September 2007.

7) Latest version available at http://www.w3.org/TR/xml11. Accessed 20 September 2007.

Index (BS 8723-4)

BSI – British Standards Institution

BSI is the independent national body responsible for preparing British Standards. It presents the UK view on standards in Europe and at the international level. It is incorporated by Royal Charter.

Revisions

British Standards are updated by amendment or revision. Users of British Standards should make sure that they possess the latest amendments or editions.

It is the constant aim of BSI to improve the quality of our products and services. We would be grateful if anyone finding an inaccuracy or ambiguity while using this British Standard would inform the Secretary of the technical committee responsible, the identity of which can be found on the inside front cover.
Tel: +44 (0)20 8996 9000. Fax: +44 (0)20 8996 7400.

BSI offers members an individual updating service called PLUS which ensures that subscribers automatically receive the latest editions of standards.

Buying standards

Orders for all BSI, international and foreign standards publications should be addressed to Customer Services. Tel: +44 (0)20 8996 9001.
Fax: +44 (0)20 8996 7001. Email: orders@bsi-global.com. Standards are also available from the BSI website at http://www.bsi-global.com.

In response to orders for international standards, it is BSI policy to supply the BSI implementation of those that have been published as British Standards, unless otherwise requested.

Information on standards

BSI provides a wide range of information on national, European and international standards through its Library and its Technical Help to Exporters Service. Various BSI electronic information services are also available which give details on all its products and services. Contact the Information Centre. Tel: +44 (0)20 8996 7111.
Fax: +44 (0)20 8996 7048. Email: info@bsi-global.com.

Subscribing members of BSI are kept up to date with standards developments and receive substantial discounts on the purchase price of standards. For details of these and other benefits contact Membership Administration. Tel: +44 (0)20 8996 7002.
Fax: +44 (0)20 8996 7001. Email: membership@bsi-global.com.

Information regarding online access to British Standards via British Standards Online can be found at http://www.bsi-global.com/bsonline.

Further information about BSI is available on the BSI website at http://www.bsi-global.com.

Copyright

389 Chiswick High Road
London
W4 4AL