

BRITISH STANDARD

Structured vocabularies for information retrieval – Guide

Part 4: Interoperability between vocabularies

ICS 01.140.20



NO COPYING WITHOUT BSI PERMISSION EXCEPT AS PERMITTED BY COPYRIGHT LAW

Publishing and copyright information

The BSI copyright notice displayed in this document indicates when the document was last issued.

© BSI 2008

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

Foreword *ii*

Introduction *1*

1	Scope	<i>2</i>
2	Normative references	<i>2</i>
3	Definitions, symbols, abbreviations and other conventions	<i>2</i>
3.1	Definitions	<i>2</i>
3.2	Symbols, abbreviations and other conventions	<i>4</i>
4	Structural models for interoperability across vocabularies	<i>5</i>
4.1	General	<i>5</i>
4.2	Model 1: Structural unity	<i>6</i>
4.3	Model 2: Non-equivalent pairs	<i>6</i>
4.4	Model 3: Backbone structure	<i>7</i>
4.5	Application of different scenarios	<i>8</i>
5	Mapping applications in context	<i>9</i>
5.1	General	<i>9</i>
5.2	The effect of different vocabulary types	<i>9</i>
5.3	Establishing mappings for index terms	<i>10</i>
5.4	Mapping of search terms	<i>11</i>
5.5	Combinations of terms	<i>12</i>
5.6	Treatment of pre-coordinated strings	<i>13</i>
5.7	Automatic mapping versus human mediation	<i>15</i>
6	Relationships and mappings across vocabularies and languages	<i>16</i>
6.1	Types of relationship	<i>16</i>
6.2	Degrees of equivalence	<i>16</i>
6.3	Equivalence between languages within a multilingual thesaurus	<i>18</i>
6.4	Mappings across structurally different vocabularies	<i>19</i>
7	Establishing equivalence for structurally different vocabularies	<i>22</i>
7.1	General	<i>22</i>
7.2	Accepting a near-match	<i>22</i>
7.3	One-to-many cross-vocabulary equivalence	<i>22</i>
7.4	Many-to-one mappings	<i>23</i>
8	Establishing equivalence between languages in a multilingual thesaurus	<i>24</i>
8.1	General	<i>24</i>
8.2	Accepting a near-match	<i>25</i>
8.3	Loan terms	<i>25</i>
8.4	Coined terms	<i>26</i>
9	Managing mappings and other relationship data	<i>27</i>
9.1	Management within one system	<i>27</i>
9.2	Management within two or more separate systems	<i>27</i>
9.3	Management external to the source and target vocabulary systems	<i>27</i>
10	Display of mapped vocabularies	<i>28</i>
10.1	General	<i>28</i>
10.2	Single record display	<i>28</i>
10.3	Multilingual thesaurus displays	<i>29</i>
10.4	Displays of mappings between structurally different vocabularies	<i>35</i>
10.5	Language and character encoding issues	<i>36</i>

11	Mapping system functionality	39
11.1	General	39
11.2	Switching or augmentation of index terms, notations or captions	39
11.3	Switching of search terms	39
11.4	Expansion of search terms	40
12	Management of projects for mapping vocabularies and languages	40
12.1	General	40
12.2	Structural considerations	40
12.3	Resources for multilingual projects	41

Annexes

Annex A (informative) Practical examples encountered during preparation of a multilingual thesaurus 42

Bibliography 51

Index (BS 8723-4) 53

List of figures

Figure 1 – Model 2 (non-equivalent pairs) as applied to four vocabularies 7

Figure 2 – Model 3 (backbone model) as applied to four vocabularies 7

Figure 3 – Single record display in a bilingual thesaurus (English – Spanish) 28

Figure 4 – Single record display in a bilingual thesaurus (Spanish – English) 29

Figure 5 – Alphabetical display for a bilingual thesaurus 30

Figure 6 – Alternative layout of entries in an alphabetical display 31

Figure 7 – Hierarchical display for a bilingual thesaurus 32

Figure 8 – Correspondence table for a bilingual thesaurus (English – Spanish) 33

Figure 9 – Correspondence table for a bilingual thesaurus (Spanish – English) 34

Figure 10 – Alphabetical display with mappings to two other vocabularies 35

List of tables

Table 1 – Tags and their equivalents in other languages 5

Table 2 – Additional abbreviations and symbols used in mappings 5

Table 3 – Elements used to represent concepts 10

Table 4 – Mapping of index terms 11

Table 5 – Mapping of search terms 12

Table 6 – Two different styles of mapping (mappings from one source vocabulary to two target vocabularies are shown) 20

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 <i>part of the scope</i> 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. <i>EXAMPLE</i> <table> <tr> <th>Source vocabulary</th><th>Target vocabulary</th></tr> <tr> <td>livestock</td><td>sheep cattle pigs poultry</td></tr> </table>	Source vocabulary	Target vocabulary	livestock	sheep cattle pigs poultry
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 <i>an aspect</i> 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. <i>EXAMPLE</i> <table> <tr> <th>Source vocabulary</th><th>Target vocabulary</th></tr> <tr> <td>women executives</td><td>women + executives</td></tr> </table>	Source vocabulary	Target vocabulary	women executives	women + executives
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.

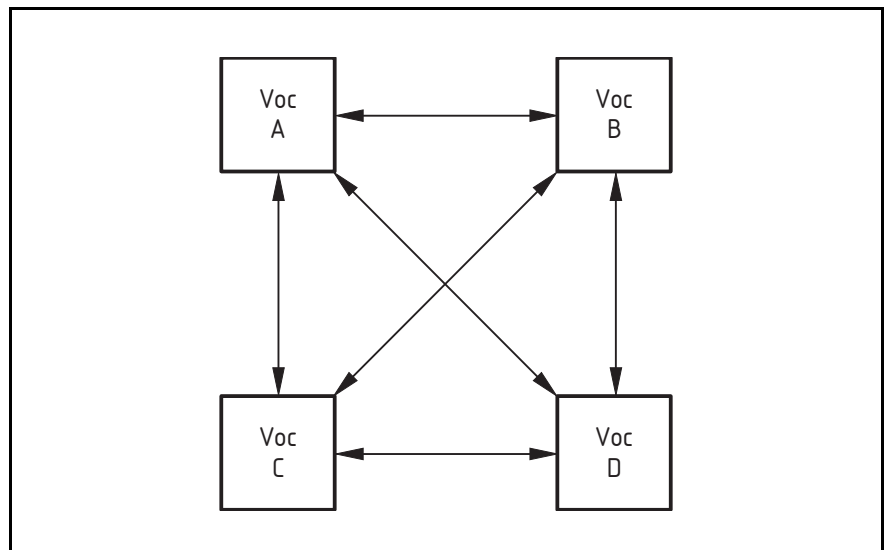
In a multilingual thesaurus, it is usual to establish a direct bi-directional equivalence relationship between corresponding preferred terms in each of the languages. (See examples in 6.3.1.) The presence of a notation often simplifies data management (see 9.1).

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.

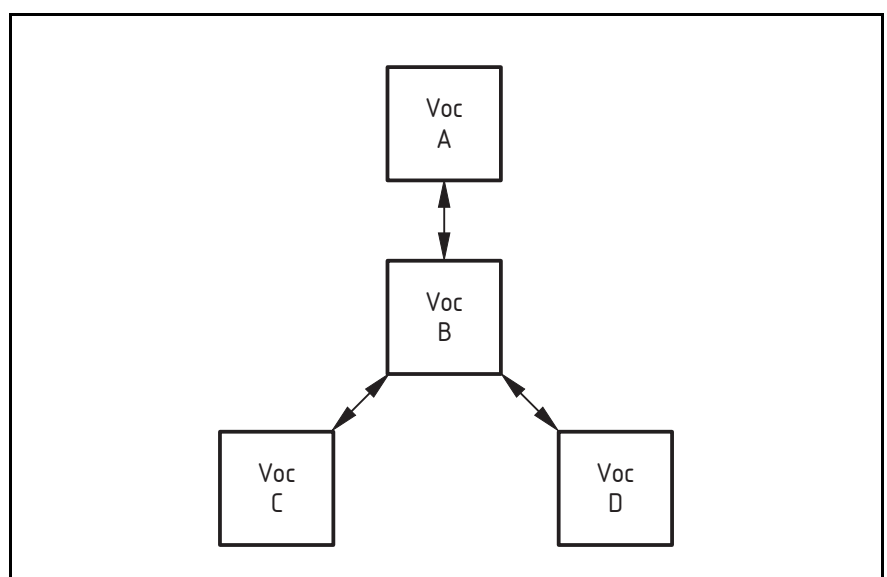
This model may be extended to any number of vocabularies/collections, by establishing direct mappings from each vocabulary to each other one. All of the collections can then be searched, using any one of the vocabularies. Figure 1 illustrates the six sets of mappings that are needed to handle four vocabularies, 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.