

BRITISH STANDARD

Structured vocabularies for information retrieval – Guide

Part 3: Vocabularies other than thesauri

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 63072 9

The following BSI references relate to the work on this standard:

Committee reference IDT/2/2

Draft for comment 07/30094121 DC

Publication history

First published December 2007

Amendments issued since publication

Amd. no.	Date	Text affected
C1	30 May 2008	Corrections to Figures 1 and 3. Corrections to formatting in 9.5.2 , Example 1

Contents

Foreword *iii*

Introduction *1*

1	Scope	<i>2</i>
2	Normative references	<i>2</i>
3	Definitions	<i>3</i>
4	Classification schemes	<i>6</i>
4.1	Vocabulary control in classification schemes	<i>6</i>
4.2	Types of classification schemes	<i>6</i>
4.3	Captions	<i>7</i>
4.4	Selection and arrangement of classes	<i>8</i>
4.5	Hierarchies in classification schemes	<i>10</i>
4.6	Notation	<i>11</i>
4.7	Alphabetical indexes to classification schemes	<i>13</i>
4.8	Presentation	<i>14</i>
4.9	Management aspects of classification schemes	<i>14</i>
5	Business classification schemes for records management	<i>15</i>
5.2	General	<i>15</i>
5.3	Hierarchy and structure	<i>15</i>
5.4	Enabling access by subject	<i>15</i>
6	Taxonomies	<i>16</i>
6.1	General	<i>16</i>
6.2	Vocabulary control in taxonomies	<i>17</i>
6.3	Choice of category labels	<i>18</i>
6.4	Provision of lead-in terms	<i>19</i>
6.5	Precise specification of scope	<i>20</i>
6.6	Notation	<i>21</i>
6.7	Hierarchical structure	<i>21</i>
6.8	Presentation	<i>24</i>
6.9	Electronic functions of taxonomies in information retrieval	<i>26</i>
6.10	Management aspects of taxonomies	<i>27</i>
7	Subject heading schemes	<i>27</i>
7.1	General	<i>27</i>
7.2	Headings and subdivisions	<i>27</i>
7.3	New schemes	<i>29</i>
8	Ontologies	<i>30</i>
8.1	General	<i>30</i>
8.2	Vocabulary control in ontologies	<i>31</i>
8.3	Choice of concepts and terms	<i>31</i>
8.4	Relationships	<i>32</i>
8.5	Presentation of ontologies	<i>32</i>
9	Name authority lists	<i>33</i>
9.1	General	<i>33</i>
9.2	Vocabulary control in name authority lists	<i>34</i>
9.3	Content of a name authority list	<i>34</i>
9.4	Authority names – selection, form and supporting notes	<i>34</i>
9.5	Relationships within an authority list	<i>35</i>
9.6	Presentation	<i>38</i>
9.7	Management aspects of authority lists	<i>39</i>

Bibliography *41*

Index (BS 8723-3) *43*

List of figures

- Figure 1 – Classification sequence with organism having priority over process 9
- Figure 2 – Classification sequence with process having priority over organism 10
- Figure 3 – From Figure 2, with a simple notation and modified chain index added 12
- Figure 4 – Example of a classification (UDC) with a notation that uses numbers and symbols 12
- Figure 5 – Example of monohierarchical taxonomy, allowing repetition of category labels with differing scopes 18
- Figure 6 – Fragments of a polyhierarchical taxonomy 22
- Figure 7 – Example of parametric searching 26
- Figure 8 – Extract from an ontology of pesticides and their uses 31
- Figure 9 – Extract from an authority list of geographical names 38
- Figure 10 – Extract from organizations section of an authority list in the education sector 39

Summary of pages

This document comprises a front cover, an inside front cover, pages i to iv, pages 1 to 44, an inside back cover 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 these committees can be obtained on request to their secretary.

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 provides 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 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.

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

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.

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

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

The ability to identify and locate relevant information among vast collections and other resources is a major and pressing challenge today. Several different types of vocabulary are in use for this purpose. Some of the most widely used vocabularies were designed a hundred years ago and have been evolving steadily. A different generation of vocabularies is now emerging, designed to exploit the electronic media more effectively. A good understanding of the previous generation is still essential for effective access to collections indexed with them.

An important object of BS 8723 as a whole is to support data exchange and other forms of interoperability in circumstances in which more than one structured vocabulary is applied within one retrieval system or network. Sometimes one vocabulary has to be mapped to another, and it is important to understand both the potential and the limitations of such mappings. In other systems, a thesaurus is mapped to a classification scheme, or an ontology to a thesaurus. Comprehensive interoperability needs to cover the whole range of vocabulary types, whether young or old.

Concepts in different vocabularies are related only in that they have the same or similar meaning. However, the meaning can be found in a number of different aspects within each particular type of structured vocabulary:

- within terms or captions selected in different languages;
- in the notation assigned indicating a place within a larger hierarchy;
- in the definition, scope notes, history notes and other notes that explain the significance of that concept; and
- in explicit relationships to other concepts or entities within the same vocabulary.

In order to create mappings from one structured vocabulary to another it is first necessary to understand, within the context of each different type of structured vocabulary, the significance and relative importance of each of the different elements in defining the meaning of that particular concept.

It follows that before BS 8723-4 can address the detail of interoperability, certain characteristics of the different types of vocabulary need to be described. BS 8723-2 describes the key characteristics of thesauri along with additional advice on best practice. BS 8723-3 focuses on other types of vocabulary and does not attempt to cover all aspects of good practice. It concentrates on those aspects which need to be understood if one of the vocabularies is to work effectively alongside one or more of the others. Recognizing that a new standard cannot be applied to some existing vocabularies, this part of BS 8723 provides informative description alongside the recommendations, the aim of which is to enable users and system developers to interpret and implement the existing vocabularies effectively.

1 Scope

This part of BS 8723 gives recommendations and guidance on some types of vocabulary, other than thesauri, that are used for information retrieval. It describes classification schemes (with business classification schemes as a special case), taxonomies, subject heading schemes, ontologies and name authority lists.

It does not give comprehensive guidance on vocabulary construction but focuses on the most common features of the vocabulary types. Where the principal application of the vocabulary is information retrieval (as is the case for classification schemes, subject heading schemes and taxonomies), recommendations are given for content, structure and some additional aspects. Where the principal application is for another function, but the vocabulary is sometimes also applied to information retrieval (as is the case with many ontologies), characteristic features of the vocabulary are described, and the only recommendations concern aspects that are useful, or which should be avoided, for purposes of information retrieval.

As authority lists of names are often applied to information retrieval, but not always to retrieval by subject, this part of BS 8723 does not give comprehensive recommendations for such lists, but describes some characteristic features.

While information retrieval is the context, this part of BS 8723 is concerned only with the vocabularies themselves and does not include recommendations for indexes, catalogues, directories or other information retrieval tools.

NOTE Thesauri and classification schemes provide two basic models, respectively, for the organization of vocabularies useful in retrieval. As their features are employed in most of the other types of vocabulary, comparisons and contrasts with thesauri and classification schemes are provided among the descriptions of the other types in order to assist with interpretation of BS 8723-4.

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, *Structured vocabularies for information retrieval – Guide – Part 1: General*

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

3 Definitions

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

3.1 caption

statement of the subjects represented by a *notation* in a *classification scheme*

NOTE A caption is read in conjunction with its hierarchical context. It does not need to be as complete or as self-contained as a scope note or even a preferred term in a thesaurus. Captions are sometimes known as class headings or class names.

3.2 category

<in a taxonomy> group of *concepts* whose members share a common feature

NOTE See also class (3.6). Not to be confused with “fundamental categories”, discussed in BS 8723-2:2005, Clause 9, in the context of faceted classification.

3.3 category label

form of words representing a *category* in a *taxonomy*

NOTE The category label in a taxonomy is comparable to the caption in a classification scheme, and although often referred to as a node, should not be confused with node label (see definition in BS 8723-1).

3.4 chain index

index to a *classification scheme*, in which entries are generated by successive left truncation of strings of terms representing *complex concepts*

NOTE See example in 4.6.2, Figure 3.

3.5 citation order

order in which *preferred terms* or *notations* are combined in a *pre-coordinate indexing* system or a *classification scheme* to form strings representing *complex concepts*

3.6 class

group of *concepts* whose members share a common feature

NOTE Classes are the basic units of which a classification scheme is constructed. In a taxonomy, although this is a type of classification scheme, they are generally known as categories (see 3.2). Occasionally they are described as “nodes”, although they should not be confused with node labels.

3.7 encoding scheme

scheme or set of rules to aid the selection and/or interpretation of values to be used in *metadata*

NOTE This term is used particularly in the context of applying Dublin Core Metadata Initiative (DCMI) standards. DCMI considers two main types of encoding scheme: vocabulary encoding schemes and syntax encoding schemes. The former type includes controlled vocabularies such as thesauri and classification schemes, whereas syntax encoding schemes provide rules for formatting or parsing of strings and are not discussed in BS 8723. Of particular note among the DCMI standards is the Dublin core metadata element set, version 1.1 [1].

3.8 identifier

set of symbols, usually alphanumeric, designating a *concept* or a *term* or another entity for purposes of unique identification within a determined context or resource, in a computer system or network

NOTE A notation may sometimes be used as an identifier.

3.9 lead-in term

term provided in a *controlled vocabulary*, not for direct use in *metadata*, but for the purpose of guiding the user to another *term* which may be used as a *category label*, *subject heading* or *preferred term*

NOTE Lead-in terms occurring in a thesaurus are generally known as non-preferred terms.

3.10 metadata

data that identify attributes of a *document* typically used to support functions such as location, discovery, documentation, evaluation and/or selection

NOTE Preferred terms or notations selected during the indexing process are commonly applied as metadata. Indexing and other functions may be carried out by human users, or by their (human or automated) agents.

3.11 name authority list

controlled vocabulary for use in naming particular entities consistently

NOTE The entities in question are unique individuals, such as Benjamin Disraeli, Kilimanjaro or the Bayeux Tapestry, rather than classes such as politicians, mountains or embroideries. A name authority list may also be known as a name authority file. In the text, a name authority list is sometimes referred to simply as an authority list.

3.12 notation

set of symbols representing a *concept* in a controlled vocabulary, especially a classification scheme

EXAMPLES

Notation	Source vocabulary	Concept
622.342 2	Dewey Decimal Classification	gold mining
373.2:372.47	Universal Decimal Classification	teaching of maths in primary schools
SBS XEJ B	Bliss Bibliographic Classification	endangered species law

NOTE 1 Notation may be used to sort and/or locate concepts in a pre-determined systematic order and, optionally, to display how the components of complex concepts have been structured and grouped. A notation can provide the link between alphabetical and systematic lists in a thesaurus. Notations are sometimes known as classmarks, class codes or class numbers.

NOTE 2 This definition supersedes that given for the same term in BS 8723-1:2005.

3.13 notation system

set of symbols, with rules for combining them to create *notations* for *concepts*

NOTE This set may be any selection of numerals, upper and lower case alphabetic characters, and punctuation symbols. The larger the set of symbols on which the notation is based, the greater the number of concepts that can be represented by distinct notations of the same length.

3.14 ontology

explicit specification of *concepts* and the relationships between them

NOTE The relationships as well as the concepts need to be defined within the particular ontology.

3.15 subject heading

term or pre-coordinated string of terms, drawn from a *controlled vocabulary*, that may be used to index a *document*

3.16 subject heading scheme

controlled vocabulary comprising single *terms* available for subject *indexing*, plus rules for combining the single *terms* in strings

NOTE Sometimes known as *subject heading language*, *subject heading list* or *SHL*. This definition supersedes that given in BS 8723-1:2005 for subject heading list.

3.17 structured heading

subject heading formed by pre-coordination of more than one *term* into a string

EXAMPLE

The single *terms* Buddhists, Mass media and Religious aspects may be combined to form the structured heading Mass media -- Religious aspects -- Buddhists

3.18 synonym ring

set of synonymous or almost synonymous terms, any of which may be used in normal text to refer to a particular concept

EXAMPLE

stream; river; brook; beck; burn

NOTE See further explanation at 6.4.2.

3.19 taxonomy

structured vocabulary using classificatory principles as well as thesaural features, designed as a navigation tool for use with electronic media

NOTE The term “taxonomy” is used elsewhere with several different meanings, but this part of BS 8723 focuses on the above definition for the sake of clarity.

4 Classification schemes

COMMENTARY ON CLAUSE 4

This standard does not give comprehensive guidance on the construction of classification schemes. The intention of this clause is to discuss issues that have a bearing on interoperability with other types of vocabulary, particularly thesauri, and to give guidance on these aspects wherever possible.

4.1 Vocabulary control in classification schemes

The basic approach of classification is to organize concepts into classes. The classes should be subdivided into narrower classes (see 4.5), and these into yet narrower classes down to the level of specificity required.

Each class should have a caption (see 4.3) and sometimes a longer description to define the scope of the class. As in a thesaurus scope note, this description may list topics that are included and may provide “see also” cross-references to other classes where related topics are located. If the scheme is multilingual, corresponding descriptions should be provided in each language represented.

Each class should also have a notation (see 4.6), typically a short string of alphabetical and/or numerical characters and/or other symbols. The notation should be independent of the language(s) of the scheme, and may be used as an unambiguous label to represent the scope of the class. The notation also serves as a convenient key for arranging items and/or locating them in the sequence of the classification scheme. Whereas a thesaurus exercises vocabulary control by designating a preferred term for each concept, a classification scheme gives a unique notation for each class.

Classification schemes should be supported by an alphabetical index (see 4.7). The index should show the location or locations of concepts in the systematic sequence, and provide extra entry points to the classes, from alternative terms and ways of expressing the corresponding concepts. The index should enable natural language access to the notation codes which constitute the artificial controlled language of the scheme.

4.2 Types of classification schemes

4.2.1 Enumerative versus synthetic schemes

All classification schemes should explicitly set out a number of classes in a schedule or set of schedules. The process of setting them out is often known as enumeration. A scheme that explicitly lays out all the classes available to the user is known as an enumerative classification scheme.

A practical problem arises when the scheme tries to provide for combinations of concepts that belong in different classes. A combined subject which has not yet been enumerated in the scheme could now need to be included, and usually it is located under one of its constituent classes, according to rules adopted by the scheme’s compiler.

As an alternative to detailed enumeration, some schemes include rules for combining concepts so that the user can synthesize compound notations from the more elementary notations of the constituent concepts. The scheme can therefore provide for very many more subjects than it is practical to enumerate. Schemes with this capability are known as synthetic or analytico-synthetic schemes.

If the notation is regarded as constituting the basic controlled vocabulary of a classification scheme, in an enumerative scheme the vocabulary is finite and explicitly laid out. For a synthetic scheme, the vocabulary is not so tightly bounded. However, the rules of synthesis limit the style and structure of the synthesized notations. Usually it is possible to decompose these into the constituent notations enumerated in the schedules.

Many large existing classification schemes today incorporate a mix of enumerative and synthetic features.

4.2.2 Faceted classification schemes

Faceted classification schemes take the principles of analysis and synthesis one step further. Firstly subjects are analysed into simple concepts according to fundamental categories such as activities, entities, places, etc. The concepts are enumerated in the scheme as classes, each with an assigned notation. Then the notation for a complex subject should be synthesized by combining the notation of the simpler concepts in accordance with rules for sequence, known as citation order. The sequencing rules are essential to ensure that all the documents on the same complex subject are given exactly the same notation and hence are brought together.

4.2.3 Implications for developers and users

Developers of classification schemes should plan any use of synthesis and/or facet analysis, in the light of 4.2.1 and 4.2.2, from the outset.

Users and implementers of an existing classification scheme should study its use of synthesis and/or facet analysis in order to interpret the synthesized notations.

4.3 Captions

As well as having a notation, each class should have a caption (sometimes called a class heading) in each language of the scheme. The caption should briefly convey the scope of the class in an unambiguous way. In most classification schemes, however, the caption does not have to be unique because the user is expected to view it in the context of the superordinate class. For example, the classes *fungi* and *fruits* might both be divided into subclasses *poisonous* and *edible*. The user is expected to understand that the concept behind *poisonous* is really *poisonous fungi* in the first case and *poisonous fruits* in the other. There is assumed to be no confusion in the location of documents on these topics, because they are given different notations. In this respect classification schemes differ from thesauri, in which the concepts are uniquely identified by terms rather than by notation.

COMMENTARY ON 4.3

The use of captions that rely on hierarchical context to convey the full meaning of the class is regarded by some as a weakness, because such captions can be misinterpreted. But it is a long-established convention in many classification schemes, and indeed in some taxonomies.

4.4 Selection and arrangement of classes

4.4.1 Aims

Two fundamental aims underlie the construction of an orderly classification scheme, namely exhaustivity and mutual exclusivity. Both aspects are difficult to achieve and compromises are sometimes needed.

Exhaustivity concerns the need to provide for every subject that is liable to occur. At the top level of the scheme, and then successively as each class is subdivided, the array should provide for every subject within the class in question.

NOTE For this purpose the parent class as well as the array beneath it may be taken into account.

The aim of mutual exclusivity is that the members of any array should be mutually exclusive, and hence reduce ambiguity in allocating a subject to the appropriate class.

While seeking to build the classification schedules in accordance with these aims, developers should take into account the established conventions described in 4.4.2 to 4.4.4.

4.4.2 The upper levels

For classification schemes with a broad scope, it is customary to divide the universe first by subject area or discipline, giving classes such as Science, Arts and Humanities, etc. The subdivisions of these are often still very broad, for example, Physics, Chemistry, Biology, etc. As a general rule, the order of top-level categories should be such that the more abstract and fundamental concepts are first, e.g. philosophy, mathematics and pure science, followed by increasingly complex concepts such as the applied sciences, technology and the arts. Some general concepts which may be combined with others throughout the schedules, such as form, time and place, are often listed in auxiliary or supplementary schedules, rather than in any discipline class.

4.4.3 Lower levels of specificity

Within each class all the concepts that are conventionally or conveniently regarded as belonging to that class are enumerated. For example, Newton's Laws may be classified under physics and chemical elements under chemistry. Similarly, general or multi-disciplinary treatments of lions may be assigned within the biology class, while a discussion of the symbolism of lions in heraldry would be more appropriately classed under heraldry. It is customary to enumerate general topics before special aspects. The extent of subdivision and the level of detail enumerated should depend on the purpose of the classification scheme and the material to which it is to be applied, so that the number of documents assigned to each class is suitable for human scanning and selection.

4.4.4 Arrangement in faceted classification schemes

In a faceted classification scheme, it is possible to apply fundamental facets (such as entities, activities, properties, etc.) at the top level. More commonly, the top level classes are based on disciplines and facet analysis is applied separately within each discipline.

In a well-constructed faceted scheme, a logical and consistent order should be used both for the order of facets and for the order of concepts within each facet. The usual order of enumeration is general before specific, i.e. general treatments of a subject are listed before specific aspects. A sequence that has generally been found suitable, which takes account both of the facets to which concepts belong and the role of concepts in relation to other concepts in the string, is:

time – space – agent – by-product – product – patient (i.e. system operated on) – operation – process – material – property – part – kind – thing

Following this order, people concerned with physics in general, such as physicists, are considered as agents, and are listed before particular kinds of physics, such as nuclear physics. The sequence may be repeated at different levels, so that nuclear physicists are listed under nuclear physics, before processes of nuclear physics such as nuclear reactions.

The order of concepts within a discipline is determined by the primary object of study. In biology, for example, a decision should be made as to whether the primary focus is organisms or processes. This implies a choice between the hypothetical examples shown in Figure 1 and Figure 2, in which, for simplicity, only extracts from those two facets have been shown.

Figure 1 **Classification sequence with organism having priority over process**

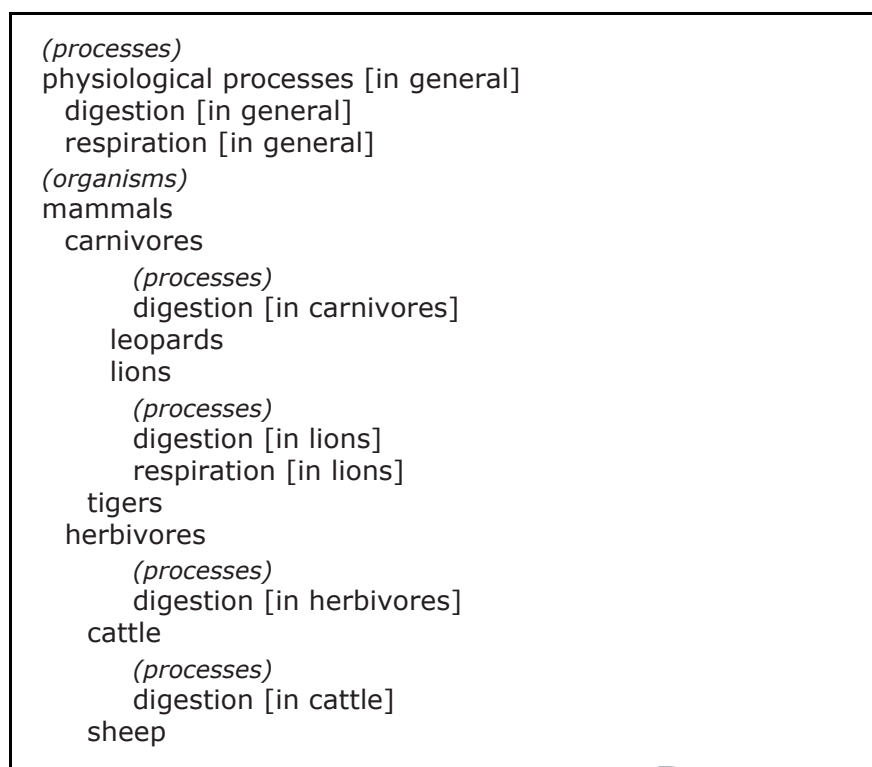


Figure 2 **Classification sequence with process having priority over organism**

(organisms)
mammals [in general]
carnivores [in general]
leopards
lions
tigers
herbivores [in general]
cattle
sheep
(processes)
physiological processes [in general]
digestion [in general]
(organisms)
[digestion in] carnivores
[digestion in] lions
[digestion in] herbivores
[digestion in] cattle
[digestion in] sheep
respiration [in general]
(organisms)
[respiration in] carnivores
[respiration in] lions

The words in square brackets in these examples are often omitted in classification schedules, being implied by the context and by indentation or typography.

In Figure 1, everything about lions is collected together, including items dealing specifically with their digestive and respiratory processes. Someone interested in all aspects of digestion needs to look for works on this topic under each species of organism.

In Figure 2, everything about digestion is collected together, including items dealing specifically with digestion in particular species. Someone interested in all aspects of lions needs to look for works on this topic under each physiological process. For the same reason, when the primary facet is *discipline*, items about particular organisms might be widely distributed; sheep and cattle might appear under both biology and agriculture, for example, and aspects of respiration might be under biology, medicine, singing and sport.

4.5 Hierarchies in classification schemes

COMMENTARY ON 4.5

This subclause gives general background information on the nature of hierarchies, explaining the differences in how these are handled, between classification schemes and thesauri.

The approach of dividing classes into narrower classes, and these into yet narrower classes, is fundamentally hierarchical. The hierarchical approach is often accentuated in the layout of a classification scheme, the size and style of typeface varying as the level of subdivision proceeds, and indentation used to show which level has been reached.