*Legal Information Management*, 4 (2004), pp. 125–130
© BIALL Printed in the United Kingdom DOI: 10.1017/S1472669604001458

# Taxonomy: Classification by Another Name

Leonard Will provides an explanation of the theoretical background to the currently fashionable taxonomy

## Taxonomies, thesauri and classifications

*Leonard Will*

The word "taxonomy" is widely used these days in discussions about organising knowledge or information, especially information in electronic form such as that on the Internet. There is no generally accepted definition of what taxonomies are, though the word is usually used to refer to alphabetico-classed schemes of subject headings. These are built on fundamental principles, which have been more fully developed in their application to thesauri and classifications. By recognising that all these kinds of "controlled vocabularies" are just variations on a theme, built from the same components, it is possible to create consistent and complementary tools to provide for the different approaches which users may need.

## Concepts as building blocks

All schemes of organisation are built from building blocks that we call "concepts". A concept is a single idea, an object, an action, or practically anything that can be expressed by a noun, possibly qualified by one or more adjectives. Examples of concepts are:

- freedom
- human rights
- insurance
- judges
- laws
- litigation
- parking tickets
- speeding
- teapots
- wigs

When we answer an enquiry, we try to find information resources dealing with the same concepts as those in the enquiry. We have a better chance of doing this if the concepts are clearly defined and consistently labelled, and this is the purpose of a controlled vocabulary.

The meanings of many concepts often appear to be self-evident from the words used to identify them, but for information retrieval purposes they may need to be more closely defined. Are "parking tickets" documents that show that you have paid for your parking or documents that show that you have parked illegally? Are "judges" met only in courts of law, or does the concept include judges of dog shows? Or is it a book in the Old Testament?

In the artificial language that is a controlled vocabulary we often define concepts by providing a "scope note" for each, and giving each concept a distinct label to identify it, called a "descriptor". To resolve the ambiguity of "parking tickets", for example, we might use the two descriptors "parking receipts" and "parking offence notifications" for the two distinct concepts.

## Finding concepts

Once we have a list of useful concepts covering the subject area we deal with, we then have to help people to find these concepts in a collection of information resources. There are two ways of doing this, conventionally called "searching" and "browsing". If someone has a specific concept in mind, then they are likely to approach the system with some words that they use to label that concept. They may search by using a single word or several words in combination. This is like using the gazetteer section at the end of an atlas to find a particular place; having found it, they can then go to the maps section and see that place in the context of its surroundings and neighbouring places. On the other hand, if the enquirer does not know any particular name, they can approach

https://doi.org/10.1017/S1472669604001458 Published online by Cambridge University Press

the problem in a "top-down" fashion, going first to a map that covers the general area and then browsing around to see what places they recognise and how they relate to one another. Thesauri and classification schemes are different ways of constructing these maps of "concept space".

## Relationships between concepts

In a thesaurus, three types of relationship are shown: equivalence, hierarchical and associative. Although these are traditionally discussed as relationships between terms, modern work on the computer representation of controlled vocabularies (e.g. Miles 2004) emphasises that the relationship is between concepts, irrespective of the terms used to label them. These relationships lead to two complementary ways of grouping concepts: by facets or by subject areas.

## Equivalence relationship

This is not really a relationship between concepts in the thesaurus, but rather an indication of alternative terms that may be used to label a single concept. A thesaurus editor may decide that it is not helpful to try to distinguish between "Acts of Parliament" and "statutes", and that they can both be included in a single concept. Either of these terms could be chosen as the descriptor for this concept, and the other one is then listed as a "non-descriptor" or "non-preferred term", with references between them, for example:

| | | |
|---|---|---|
| *Acts of Parliament* | USE | **statutes** |
| **statutes** | USE FOR | *Acts of Parliament* |

There is no implication that one term is necessarily more "correct" than the other, though it is usual to choose as a descriptor the term that is most likely to be used by users of the system and that is most likely to convey the meaning of the concept as it is defined in the scope note. A good search system should ensure that a user is led to the concept whichever term is used in the enquiry. It is often helpful to distinguish descriptors and non-descriptors typographically, for example by using bold and italic, as has been done in this article.

## Hierarchical relationship

Concepts can be linked together into trees by recognising that some concepts are subsets of others. For example **barristers** and **solicitors** are both "kinds of" **lawyers**, and **lawyers** and **legal assistants** are "kinds of" **legal personnel**. This is sometimes called the "is-a" relationship, e.g. a **solicitor** "is-a" **lawyer**. The conventional way of representing this in a thesaurus is to use the symbols BT and NT, standing for "broader term" and "narrower term", e.g.

**barristers**
BT **lawyers**

**lawyers**
BT **legal personnel**
NT **barristers**
**solicitors**

**legal assistants**
BT **legal personnel**

**legal personnel**
NT **lawyers**
**legal assistants**

**solicitors**
BT **lawyers**

This relationship should be defined only when it is true irrespective of context. For example, the relationship

**wigs** BT **legal attire**

should not generally be used. It might be thought valid in a legal thesaurus, but this is not necessarily the case, as identification evidence might be disputed on the grounds that the defendant wore a toupee at the time of an alleged crime. **Speeding** might or might not be a valid narrower term of **traffic offences** depending on whether its scope note defines it as "travelling fast" or "exceeding the speed limit".

The relationships above can also be represented as a tree structure, e.g.

```
legal personnel
    lawyers
        barristers
        solicitors
    legal assistants
```

Well designed search software will allow an enquirer to ask for "all kinds of **legal personnel**", with the system then retrieving items indexed by any of the terms below this term in the tree structure. A single concept may have more than one broader term, and thus appear in more than one place in a hierarchy, so that **teapots** could be listed as a narrower term of both **containers** and **tea preparation equipment**. A thesaurus that allows this is said to be "polyhierarchical", in distinction to a biological taxonomy, which is "monohierarchical".

## Concepts grouped into facets

If we build all concepts into hierarchies in this way, introducing higher-level terms where necessary to create valid groups, we eventually arrive at a comparatively small number of general categories that cannot be combined further. These might be concepts such as **people**, **objects**, **activities**, **abstract concepts**, **disciplines**, **places** and **times.** These general categories are called "facets". It is not possible to specify a single definitive list of facets, as there is an element of subjectivity in their choice, but

https://doi.org/10.1017/S1472669604001458 Published online by Cambridge University Press

they are normally defined so as to be mutually exclusive. A concept cannot be both a person and an activity, for example. The term "facet" is, unfortunately, used with different meanings in the professional literature, which causes much confusion. The definition given here is the one that I find most useful and agrees with the current draft revision of the British Standard for thesaurus construction, which defines a facet as a "high-level grouping of concepts of the same inherent category".

## Associative relationship

There are some concepts that are closely related, so that someone searching for one might wish to have their attention drawn to items concerning the other, but where the relationship is not hierarchical, such as that between **advocacy** and **barristers**. This is represented by the "related term" (RT) relationship, which is shown in both directions, e.g.

| | | |
|---|---|---|
| **advocacy** | RT | **barristers** |
| **barristers** | RT | **advocacy** |

Good search software will draw these related terms to the attention of searchers, allowing them to add them to their searches if they wish.

## Concepts grouped by subject area

The hierarchical relationship discussed above provides a way of grouping concepts based on their fundamental categories to which they belong, and allows for the kind of query expansion discussed in the last paragraph under "Hierarchical relationship" above. A person browsing a collection, and seeking a map that brings related topics together, will often prefer to see concepts grouped by the discipline or subject area to which they relate. If they are interested in law courts, for example, they may wish to see the judges, counsel, litigants and juries (people) grouped with rules and regulations (documents) which control their powers and rights (abstract concepts), the court buildings (places) and the apparatus and exhibits that may be used in a case (objects). Concepts from many different facets are thus brought together, and we need to find a way of doing this that arranges concepts in a logical and understandable sequence. Related topics should be brought together and unrelated topics kept separate. This is the role of classification, which will be discussed further below.

## Indexing and searching with a thesaurus

When we index a document with a thesaurus, we give it the descriptors for all the concepts it contains that are likely to be sought. These descriptors are assigned independently, and not linked together, so a document on the cross-examination of expert witnesses by counsel in criminal courts might be given the descriptors

**barristers**
**criminal courts**
**cross-examination**
**expert witnesses**

This would then be retrieved by search statements that asked for one or more of these terms. A search for **expert witnesses** or a search for **barristers** would retrieve it. It is unlikely that all the terms used in a search would exactly correspond to the terms used in indexing, but greater precision would be achieved by combining terms into more complex searches, such as

(**barristers** OR **counsel**) AND **expert witnesses**

This technique of providing for terms to be combining or coordinated after indexing (at the time of searching) is known as "post-coordinate indexing".

## Classification

The alternative technique, known as "pre-coordinate indexing" consists in combining terms at the time of indexing to express compound subjects. Instead of the four separate terms shown in the previous section, we can combine them into a string such as:

**expert witnesses : cross-examination : barristers : criminal courts**

This is useful, because a collection of strings of this kind can be arranged alphabetically, like the index at the back of a book. It brings everything about the first-cited concept together, arranged in a logical order, and the rest of the string gives a summary of the main concepts in the document to help browsing and choosing the most relevant resources.

The terms in a string like this can be combined in many different ways, and we cannot list them all; there are 24 different permutations of four terms, for example. We therefore need a rule to determine which sequence is to be used. This is sometimes referred to as a rule for the "citation order of facets", but using facets (as defined above) is not sufficient; **expert witnesses** and **barristers** are both members of the **people** facet, so that will not tell us which should come first. The rule appropriate to any subject field is a matter of judgement, but it should be applied as consistently as possible and we do this by examining the role of each element in the string.

A simple rule that has been found suitable in many cases is to combine concepts in the sequence

person or thing acting – intransitive action

e.g. **traders** : **accounting**
or **ships** : **sinking**

https://doi.org/10.1017/S1472669604001458 Published online by Cambridge University Press

person or thing acted on – transitive action – person or thing acting

e.g. **planning applications : rejection : [by] inspectors**

or **victims of crime : compensation : [by] criminals**

A frequently cited fuller form is

thing – kind – part – property – material – process – operation – patient – product - by-product – agent – space – time

where "thing" is the focus or core subject concerned and "patient" is the person or thing acted on. Only some of these elements are likely to be used in any one string of indexing terms, and some of them, such as "product" and "by-product" are mainly relevant to manufacturing and technical topics and not likely to be much used in legal indexing.

In many classification schemes, it is found that an initial grouping by *discipline* or area of work or study is desirable, before applying the sequence above. In law, two other important factors are *jurisdiction* or *system of law* and the *form* of a resource. Examples of these are

*discipline*: commerce, insurance, intellectual property, constitutional law

*jurisdiction or legal system*: international law, common law, Scots law, European Community law

*form*: statutes, cases, digests, statutory instruments, treatises, journals

I cannot specify here a definitive order for the components of an indexing string. Anyone constructing a classification or taxonomy should, however, decide on the order that is most appropriate and useful to their users and then apply it consistently. If, for example, all material relates to a single jurisdiction, then that element can be left out. If it is retained, one possible sequence, *jurisdiction – subject area (discipline) – topic within discipline – object (thing) – form*, would give a list of headings like the following:

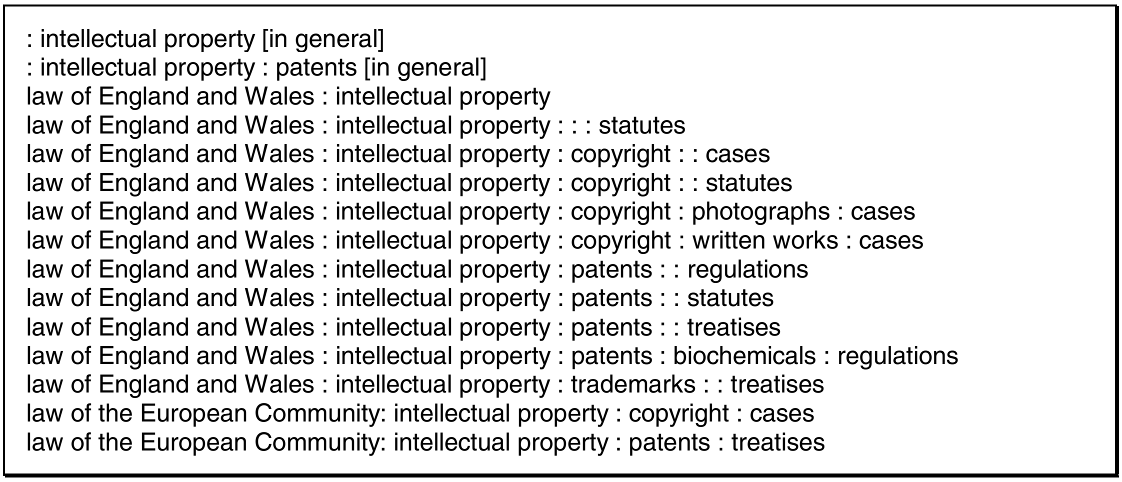
These strings do not all contain five elements, but where an element is missing the colon preceding it has been retained so as to maintain the filing order, ensuring that general works are listed before works where the same subject occurs in specific contexts.

## Symbolic notations

Rather than relying on simple punctuation, some classification systems use distinctive symbols between terms in order to maintain filing order, or label each string with a "notation" such as a classification number. This allows strings to be sorted in the most useful order, which may not be alphabetical but could, for example, be in the sequence in which events occur, or in order of increasing complexity. A notation also serves as a way of linking strings to entries in an alphabetical index of terms. It is important to realise that notation is an auxiliary device applied to a classification after the order of subjects has been determined; it is a common fallacy to believe that if, for example, classification numbers use only the ten digits from 0 to 9, this limits the number of sibling terms to ten at any level of subdivision.

A sequence of headings such as those shown in Figure 1, in a classified order arranged by the words in each heading rather than by a separate notation, is called "alphabetico-classed"; this is the usual form in which "taxonomies" are presented. Arrays of terms at the same level are in alphabetical order, though they may be organised into more systematic groups by the insertion of additional levels of headings.

The set of strings shown above could be represented as an indented display as shown in Figure 2. In this example *form* has been distinguished typographically by being printed in italics, but it is also possible, and often helpful, to include explicit "node labels" such as *(topic)* or *(form)* as in Figure 3.



: intellectual property [in general]
: intellectual property : patents [in general]
law of England and Wales : intellectual property
law of England and Wales : intellectual property : : statutes
law of England and Wales : intellectual property : copyright : : cases
law of England and Wales : intellectual property : copyright : : statutes
law of England and Wales : intellectual property : copyright : photographs : cases
law of England and Wales : intellectual property : copyright : written works : cases
law of England and Wales : intellectual property : patents : : regulations
law of England and Wales : intellectual property : patents : : statutes
law of England and Wales : intellectual property : patents : : treatises
law of England and Wales : intellectual property : patents : biochemicals : regulations
law of England and Wales : intellectual property : trademarks : : treatises
law of the European Community: intellectual property : copyright : cases
law of the European Community: intellectual property : patents : treatises

*Figure 1: Alphabetico-classed list of headings.*

https://doi.org/10.1017/S1472669604001458 Published online by Cambridge University Press

- intellectual property
    - patents
- law of England and Wales
    - intellectual property
        - *statutes*
        - copyright
            - *cases*
            - *statutes*
            - photographs
                - *cases*
            - written works
                - *cases*
        - patents
            - *regulations*
            - *statutes*
            - *treatises*
            - biochemicals
                - *regulations*
        - trademarks
            - *treatises*
- law of the European Community
    - intellectual property
        - copyright
        - patents
            - *treatises*
            - *cases*

*Figure 2: Classified display.*

A display such as Figure 2 can be interpreted as a nested series of menus. Someone navigating downwards can choose first the jurisdiction in which they are interested, then within that the subject area: intellectual property, planning law, taxation etc. If they choose intellectual property, the next level of menu might present the options of copyright, patents and trademarks, and so on. Web sites often keep track of the steps which users take to work down to a specific page, and display the path taken at the top of the page. Someone looking at a page discussing cases on the copyright of photographs might see the path

> law of England and Wales > intellectual property > copyright > photographs > cases

A path like this is sometimes called a "breadcrumb", because it shows the user the path they have taken and allows them to find their way home, like the trail of breadcrumbs left by Hansel and Gretel as they went into the woods. (Lida et al., 2003). It is just the same as one of the alphabetical subject strings shown in Figure 1.

A specific concept may occur in many different subject strings, and it will not be found in the alphabetical sequence unless it is the first element of a string. It may be made retrievable in three different ways:

- a computer may search for terms occurring in any part of a string; it can then display all the strings in which these terms occur, allowing the user to select one or more and start navigating from that point;
- an index may be generated in which each term is used in turn as the filing element, by rotating strings or creating a "chain index". "Keyword in/out of context" (KWIC/KWOC) indexes may be used if the list is to be printed, but these are little used now that most searching is done on computers;
- additional sequences of subject headings may be provided, like Figure 1, but using a different citation order for the elements that make up the string. If this is done, the different sequences should be kept as separate lists, so that the citation order within each list is consistent.

- *(subject area)*
- intellectual property
    - *(topic in intellectual property)*
    - patents
- *(jurisdiction)*
- law of England and Wales
    - *(subject area)*
    - intellectual property
        - *(form)*
        - statutes
        - *(topic in intellectual property)*
        - copyright
            - *(form)*
            - cases
            - statutes
            - *(objects of copyright)*
            - photographs
                - *(form)*
                - cases
            - written works
                - *(form)*
                - cases
        - patents
            - *(form)*
            - regulations
            - statutes
            - treatises
            - *(objects of patents)*
            - biochemicals
                - *(form)*
                - regulations
        - trademarks
            - *(form)*
            - treatises

*Figure 3: Classified display with node labels.*

## Conclusion

Though computers can search for every document containing specific words, this is often not the most effective way to retrieve the most relevant material. It also doesn't usually organise the results in a way that makes it easy to browse through them. The use of a controlled vocabulary can help, and even automatic categorisation software can work better when it is given an intellectually constructed framework to start from. The principles of organising information have been developed over many years, and should not be forgotten just because the work is relabelled by a fashionable name such as "taxonomy".

https://doi.org/10.1017/S1472669604001458 Published online by Cambridge University Press

## References and further reading

Aitchison et al. (2000). Thesaurus construction and use: a practical manual/Jean Aitchison, Alan Gilchrist, David Bawden. – 4th ed. – London : Aslib, 2000. – 240p; 30cm. – ISBN 0-85142-446-5

British standard guide to establishment and development of monolingual thesauri/British Standards Institution. – 1st rev. – London : BSI, 1987. – 32p; 30cm. – (BS5723:1987) (ISO2788-1986). – [New edition in preparation]

Foskett (1996). The subject approach to information/A. C. Foskett. – 5th ed. – London: Facet Publishing. – 472p; 23cm. – ISBN 1-85604-048-8 1996

Lida et al. (2003) Breadcrumb navigation: an exploratory study of usage/by Bonnie Lida, Spring Hull and Katie Pilcher. – *Usability news* 5.1. <http://psychology.wichita.edu/surl/usabilitynews/51/breadcrumb.htm>

Miles (2004). RDF thesaurus/maintained by Alistair Miles. – *([Semantic Web Advanced Development] SWAD-Europe thesaurus activity)*. <http://www.w3c.rl.ac.uk/SWAD/rdfthes.html>

Leonard Will is a consultant in information management, with a special interest in subject indexing, thesauri and classification schemes. He is a member of the Classification Research Group and of a working party currently drafting a new British Standard BS8723 "Structured vocabularies for information retrieval" which will replace the existing BS guidelines on thesaurus construction. Previously Head of Information and Library Services at the Science Museum, London, since 1994 he has run his own business, "Willpower Information", in partnership with his wife, Sheena. <http://www.willpowerinfo.co.uk/>

*Legal Information Management*, 4 (2004), pp. 130–135
© BIALL Printed in the United Kingdom DOI: 10.1017/S1472669604001483

# Information Literacy: Organisational and Law Firm Perspectives

Jane Macoustra, previously of Clifford Chance's Hong Kong office and now a freelance consultant, shares with us her views on the challenges of implementing information literacy programmes within organisations

## Introduction

This article is an extended version of my previous article "Information Literacy in a corporate environment" which was published on the FreePint website last year. http//:www.freepint.com/issues/060303.htm. The opinions in this article are solely those of the author. Feedback or discussion is welcomed.

As an information professional, IL is a competency that I have taken for granted, because it is a natural part of what being an IP is all about, but other people working in a corporate organisation may very well not possess these skills, through no fault of their own. IL has been around for a long time and is a well-documented subject – especially in an academic context in Australia and the USA. There is not so much information available when it is translated across to an organisational or workplace environment. In this article I consider IL in the environment of the legal or corporate organisation to enable the reader to understand further the concepts that are involved.

*Jane Macoustra*