

# Indexing the law: a controlled vocabulary

Mark Scott

---

*Recent years have seen a proliferation of sources of information available to the lawyer. This article describes one publisher's efforts to provide a more effective means of accessing information across the range of media available. It outlines the development of a controlled vocabulary and its application to both electronic and printed indexes.*

---

## Background

Sweet & Maxwell is a leading provider of information and solutions to the legal and professional markets in the UK and Ireland. It has a history of over 200 years in legal publishing, providing legal information to lawyers, academics and students. Legal information comes in many forms, the characteristic weighty textbooks being only one. Material relevant to the legal market includes virtually all the paper emanating from government, 'primary sources' which are the texts of legislation and the transcripts of court judgments, and 'secondary sources' which include commentary, textbooks and journals. The legal professional relies on having access to this wide body of information at all times in a world where knowledge and understanding of the latest case not only provides an advantage but may also avoid a claim of professional negligence.

Alongside information sourced externally, all law firms produce a large quantity of information internally. Every working day, news bulletins, reusable forms of contract and detailed analyses of legal developments are being produced within law firms, with some of the larger firms taking the further step of publishing internally produced material commercially. The value in the information produced within law firms is perhaps best demonstrated by the relatively long history of knowledge management in the legal sector. More commonly referred to as know-how, conscious efforts in this direction have been made since the 1960s.

In common with most other areas of information provision, the most important development in recent years has been the increase in electronic sources. In particular the provision of primary materials (cases and legislation) has become largely an electronic business. Printed work however remains important – a solicitor's office would not be the same without a few leather-bound volumes on the shelves.

The provision of primary material over the Internet has also resulted in the availability of information at no charge. Nearly everything produced by the European Union is available free at <http://europa.eu.int>, new UK legislation is available at <http://www.legislation.hmso.gov.uk>, and the courts themselves provide access to an increasing volume of case law. The British and Irish Legal Information Institute brings many of these sources together at <http://www.bailii.org>. Publishers, therefore, have to place increasing emphasis on adding value to this information. A key component in this is providing assistance in effective and efficient retrieval.

## Benefiting from a controlled vocabulary

Sweet & Maxwell has identified the application of a controlled vocabulary as an important means by which information retrieval may be enhanced. The merits and possible weaknesses of controlled vocabularies have been widely discussed (for a useful summary of the relative strengths and weaknesses of natural language and controlled language see Aitchison, Gilchrist and Bawden, 2000: 5–7), but I will outline a couple of factors that I believe are of particular relevance to legal materials.

Very few of a lawyer's information needs will be met by reference to just one source. A typical query may start with reference to a textbook, followed by a check on recent case law or legislation, and finally with recourse to up-to-date commentary in a journal. The implementation of a controlled vocabulary across all of these sources allows a single query to be formulated, thereby reducing the work that has to be done by the researcher. As well as involving a number of sources, information requirements have to be met across a range of media. Although the advantages to be gained from this approach become more obscure when viewing individual sources in isolation, particularly traditional paper sources, the greatest benefits arise when the entire collection has been indexed consistently. What is created is, in effect, an encyclopaedia of English law. All of the content indexed in accordance with the controlled vocabulary becomes accessible by means of each and every concept described in the vocabulary.

One difficulty that has been associated with the use of controlled vocabularies in indexing has been the desirability of using the language used by the author of any item. This has been identified as being particularly important for legal materials.

Law and legal principles can be expressed only in words . . . Therefore vocabulary is of special importance to lawyers . . . It follows that the subject index to a law book must contain the lawyers' own terms, as they use them, with the minimum of constructive intervention from the indexer. (Moys, 1992)

The ideal situation is achieved where the language of the author and the language of a controlled vocabulary coincide. With a properly selected controlled vocabulary that will be the case more often than not. Where differences arise, it is as a result of there being more than one way of expressing a particular concept, thus returning us to one of the principal benefits of a controlled vocabulary, the control of synonyms. A controlled vocabulary can define terms precisely, and its

proper application by skilled and knowledgeable indexers will minimize the danger of the misinterpretation of any concept expressed.

## Thesaurus and taxonomy

The development of Sweet & Maxwell's Legal Taxonomy began with the creation of the Legal Journals Index by Legal Information Resources in 1986. This took the form of a rotated keyword index, and to produce it consistently a thesaurus was created (for further detail see Smith, 1987 and Miskin, 2002: 21). Following the acquisition of Legal Information Resources by Sweet & Maxwell, the Legal Thesaurus came to be used in other products. The most notable was *Current Law*, a monthly publication producing summaries of legislation and case law, and by extension its electronic equivalent Current Legal Information.

Once the potential benefits of a wider application of the Thesaurus had been identified, the decision was taken to redevelop it. This was deemed necessary to ensure that the needs of all Sweet & Maxwell's content could be met. Partly as a result of the Thesaurus's evolution to meet the needs of journal indexing, large areas of law were either not covered or not covered in sufficient detail to meet the requirements of books or primary sources. With a need to introduce so many new terms identified, a restructuring was also required. It was clear that without the adoption of a standard structure it would become increasingly difficult to navigate.

The redevelopment of the Legal Thesaurus also involved its renaming as the Legal Taxonomy to distinguish the old from the new. Although it is described as a taxonomy it does in fact contain the thesaural relationships of equivalence and association alongside the polyhierarchical taxonomic structure.

In creating any controlled vocabulary, the first thing to bear in mind is its purpose. The stated aim of providing a controlled means of indexing all of Sweet & Maxwell's content is clearly the starting point, but it is necessary to look closer than that. The company provides legal information emanating from a number of different jurisdictions, each of which has a legal system composed of different institutions and often built upon different legal concepts. The core of Sweet & Maxwell's business is English law, and it is English law that forms the core of the Legal Taxonomy. Naturally European law forms an integral part of most legal practice today, and is reflected in the Taxonomy. More difficulties arise with the need to reflect Scots and Irish law. A multilingual solution does not provide the answer. Although many of the differences between legal terms used in the different jurisdictions are purely linguistic, many are substantive, and very fine differences in meaning can make a crucial difference in the legal arena. It has therefore been necessary to enter separate terms for Scots and Irish purposes; this can lead to some difficulties with the structure of the Taxonomy which remains solidly English law-based. This is a fundamental characteristic of the Taxonomy which is not amenable to compromise.

Another fundamental characteristic of the Taxonomy is its size. The Taxonomy had to cover the subject matter of the

whole of Sweet & Maxwell's product offering, legal information in the UK and Ireland, and at a level of detail sufficient to index all of that content, as detailed as the index for the most specialist work in any particular subject area. There are many ideas about the optimum depth and breadth of a taxonomy, but with 111 top terms the Taxonomy far exceeds them all. The tendency of the legal sphere to divide itself into specialisms has brought this about. However it also serves to mitigate the impact of the exceptional size, since most users will only work within a discrete number of areas within the Taxonomy, and the only users who will have to be able to operate around the Taxonomy as a whole will be information professionals of one sort or another, whether law firm librarians, indexers working for Sweet & Maxwell or taxonomy developers.

The top terms chosen reflect the common divisions of the profession into practice areas. Many of these are obvious, and it is difficult to imagine any structured description of law that does not have headings such as Administrative law or Employment. Other decisions are less easy to make, and their selection was informed by customer research as well as the experience gained in over ten years of development of the Legal Thesaurus.

One of the most notable changes to the structure was the adoption of a polyhierarchical as opposed to a monohierarchical system. Since the same legal concepts such as 'Public interest' and 'Reasonableness' are pertinent to many practice areas, it was necessary to have terms occurring more than once in the taxonomy structure. The lack of this flexibility in the old scheme had led to many terms only being accessible to those users familiar with the structure. It was perfectly possible that a conceptually valid hierarchical route to a particular term would lead to a dead end.

Another consequence of the diverse relevance of many terms in different areas of law was the adoption of a partially faceted scheme. Such facets as 'Persons' or 'Place' are not unusual, and occur frequently where a faceted approach has been adopted. Some of the more particularly legal facets such as 'Courts' and 'Liabilities' are a feature of the particular subject matter involved. The repetition of these facets under many top terms allows for a higher level of standardization, and again serves to mitigate the impact of the division of the Taxonomy into so many top terms. With the structure below them becoming more predictable by the use of facets, a more consistent approach may be adopted by the user.

As mentioned above, equivalence and association are contained within the Taxonomy. The importance of equivalence relationships cannot be overstated. The inclusion of synonyms in controlled vocabularies performs the same function as the cross-reference in a print index, providing the user with multiple entry points and avoiding needless repetition. A useful side-effect of the diversity of the Legal Taxonomy's applications is that each different discipline informs the creation of new equivalence relationships within the Taxonomy, and they in turn inform each new application.

Associative relationships are particularly useful in the context of the semi-faceted approach that has been adopted. The level of detail and resulting number of terms have meant that any assistance to the user of the Legal Taxonomy

becomes more valuable, and navigating between facets by means of association provides such a tool, whether between persons and activities ('Childminders' and 'Childcare') or entities and their attributes ('Aircraft' and 'Airworthiness').

An aid to users is the inclusion of detailed Scope notes providing definitions and usage instructions. Particularly helpful in a legal context is the inclusion of references to statutes. The term 'Aggravated vehicle taking' refers the user to the Theft Act 1968 s.12A, a useful shorthand means of giving an exact definition.

Further details on the structure of the Legal Taxonomy are available at

[www.sweetandmaxwell.co.uk/online/taxonomy/index.html](http://www.sweetandmaxwell.co.uk/online/taxonomy/index.html).

## Application to data

The first decision to make is the level at which the information at hand is to be indexed. With such different types of document this decision needs to be taken a number of times, considering the structure and contents of all the material. The level chosen is the level at which an item has an individual relevance. It must have the potential to stand alone as the answer to a query. So an Act of Parliament is broken down to its sections, a textbook to its paragraphs, whereas a court judgment is considered as a whole.

To each individual document a keyword string is attached. This will contain as many terms as are necessary to fully describe the contents of the document. As well as keywords, one or more subject headings are assigned to documents. The subject headings correspond to the top terms of the Taxonomy, and allow much greater precision in describing concepts by the qualification of keywords. One potential problem arising from a polyhierarchical system is that, although a term may mean the same thing with each occurrence, it may imply a different context depending on where it occurs. For example a 'Stay of proceedings' is the same thing whenever it occurs, but knowing whether the proceedings that have been stayed are civil or criminal is clearly going to be useful. This can be reflected by the use of the subject heading 'Civil procedure' or 'Criminal procedure'.

Where the document is part of a print product, the keyword string may be utilized in the creation of a back of book index (see the example below). Each keyword in the string can form a main entry, with each of the other keywords providing the next-level terms as appropriate. Where necessary, a further text entry can provide extra detail. The general rule is that each main entry must be a keyword from the Legal Taxonomy and, where possible, the sub-entries will also be taxonomic keywords.

### Example

Keyword string: Duty of care; Solicitors

Index entries:

```
duty of care
  solicitors
    scope of duty
solicitors
  duty of care
    scope of duty
```

Exceptions of course have to be made. The needs of a particular text cannot be ignored, and aspects of a book are often reflected in index terms that are not, strictly speaking, dealing with the concepts expressed. A law book will have references to the titles of legislation and other such documents. Also the meaning of a particular word or phrase may be discussed which falls outside the scope of the Taxonomy. As long as the user is made aware that these exceptions exist, any impact to the wider scheme is minimized.

In electronic products the key to the utilization of any indexing is the user interface. It is vital that any interface developed for use alongside a taxonomy makes full use of the features provided by the taxonomy. Allowing searches to define keywords or subject headings and automatically accounting for synonyms and narrower terms are clearly key features of any interface. But searching is not the only application. More of the benefits of a taxonomy may be demonstrated through the use of browsing. The mechanisms by which a user may navigate around a hierarchical structure are increasingly familiar to everyone using online services, and it is the responsibility of the taxonomy itself to ensure that the routes taken through a subject are intuitive. Again it is worth considering the target audience for a taxonomy implementation. The natural route through a subject will not be the same for the specialist and the lay person. Consequently the Legal Taxonomy is designed to be navigable by persons with legal knowledge, but more than that, it is the perspective a legally trained person has of the law that must inform the shape of the taxonomy.

The beauty of electronic applications as opposed to paper-based examples is the flexibility that is available. For example, the extent to which users should be made aware of associative relationships and definitions within the Taxonomy may well vary between applications and between users. The searching, or browsing, experience can be tailored to particular data sets and particular levels of user knowledge and experience.

## Further advantages

There are advantages to be gained beyond those of adding value to the customer offering. Keywords need only be created once for each item, although some work will always be required to tailor keywords to the needs of a particular product index. Much of the effort, and therefore cost, is saved. Further, it is not only customer access to Sweet & Maxwell data that is improved. The use of a comprehensive taxonomy allows product developers to have access to the company's internal repository of documents, providing opportunities to build new products, collecting material together on the basis of what it is about. Furthermore, an additional tool is made available for updating existing products: consistent indexing eases the addition of new primary references to existing secondary sources.

As mentioned above, an increasingly important information source in the legal world is the internal documentation of law firms. The Legal Taxonomy provides a standard by which both internal and external information can be described and accessed in the same way. For many years the Legal Thesaurus was used in law firms to organize

document collections, sometimes for ordering books on shelves and in other applications as part of a knowledge management initiative. In light of this existing acceptance, the development of the Legal Taxonomy provides further opportunities for integration. The development of a more detailed scheme matches the needs of specialist law firm knowledge management schemes, where very fine levels of granularity are required. The advantages are clear for legal knowledge managers of having part of the structure of a knowledge management system looked after for them, and having a large quantity of relevant data already indexed accordingly. Very few UK law firms can call upon the internal resource to create and manage a taxonomy. Making a taxonomy available as a stand-alone piece of software or as individual files that can be integrated into a knowledge management system allows a large investment of time and effort to be avoided.

## Conclusion

In a world of ever more diverse sources of information, any technique that makes retrieval of information easier has to be welcomed. A controlled vocabulary provides one such technique, which can be applied to a wide range of sources

across a range of media. Ensuring that a controlled vocabulary contains the level of detail needed means that a lot of work is required to establish the vocabulary in the first place, but this effort is outweighed by the benefits in terms of consistency and accessibility across the whole range of data to which it is applied.

## References

- Aitchison, J., Gilchrist, A. and Bawden, D. (2000) *Thesaurus construction and use: a practical manual*, 4th edn. London: Aslib.
- Miskin, C. (2002) Taxonomies. *Legal Information Management* 2(1), 16.
- Moys, E. M. (1992) Legal vocabulary and the indexer. *The Indexer* 18(2), 75.
- Moys, E. M. (1993) *Indexing legal materials: occasional papers on indexing no. 2*. London: Society of Indexers.
- Smith, N. (1987) Indexing legal journals. *The Law Librarian* 20(1), 1.

Mark Scott is a Senior Taxonomy Analyst at Sweet & Maxwell Group. Mark has worked in legal information provision for seven years after gaining a degree in law from the University of Edinburgh and a diploma in information and library studies from the Robert Gordon University in Aberdeen. Email: mark.a.scott@thomson.com

### **Indexing the medical sciences by Doreen Blake, Michèle Clarke, Anne McCarthy and June Morrison**

Society of Indexers Occasional Papers on Indexing, No. 3. 2nd edn, 2002, 84 pp., ISBN 1-871577-24-1

£17.50 (£20.00 overseas); £15.00 (£17.50 overseas) for members of indexing societies



Guidance on the pitfalls and possibilities in medical indexing, including a good selection of interesting and specific examples. The text of the previous edition has been extensively rewritten, and advice is given on a wide range of topics, including:

- ★ medical terminology
- ★ choice of headings
- ★ journals indexing
- ★ indexing names
- ★ abbreviations and acronyms
- ★ final preparation of the index text.

There are also new sections on:

- ★ electronic indexing
- ★ the ethical aspects of changing language
- ★ gene nomenclature.

The comprehensive list of reference materials includes up-to-date website addresses.

A very practical book, with a wealth of friendly commonsense advice... much encouragement to those with less experience. (*The Indexer*)

An excellent entrée into the art of medical indexing. (*IASC Bulletin*)

The Occasional Papers on Indexing provide more detail than can be accommodated in general textbooks on indexing and are written by practitioners who bring specialist experience and expertise to their subjects. The series aims to contribute to raising standards of indexing, and to stimulate indexers to think more deeply about indexing principles and working methods.

To order this and other SI publications, visit [www.indexers.org.uk](http://www.indexers.org.uk) for a downloadable order form or contact: Sales Administrator, Society of Indexers, Blades Enterprise Centre, John Street, Sheffield S2 4SU, UK Tel: +44 (0)114 292 2350 Fax: +44 (0)114 292 2351 email: [admin@indexers.org.uk](mailto:admin@indexers.org.uk)