

9

CREATING AND MANAGING A THESAURUS FOR ACCESSING NATURAL SCIENCE COLLECTION AND OBSERVATION DATA

CHARLES J.T. COPP

A thesaurus is a key enabling component for effective data retrieval from large data networks such as will be established through *ENHSIN* and the follow-on *BioCASE* project. The *BioCASE* thesaurus will include multiple related-term lists covering all aspects of natural science collections and field observations including taxonomy, habitats, gazetteers, collecting methodologies and stratigraphy. The thesaurus will initially be populated with classifications and term lists derived from existing sources, but will be used for storing and checking words derived from collection metadata, database indexing and terms entered by users in free text queries. It will enable these terms to be placed in a meaningful context with other equivalent terms and both broader and narrower categories useful for maximising or refining the number of returns to queries. It is not the purpose of such a thesaurus to be correct, comprehensive or authoritative or to act as a terminology standard although all of these things could be true for some of its content.

The *BioCASE* thesaurus will be very large in size, containing several million terms and initial tests show that indexing will lead to a rapid rise in the number of new terms added to the thesaurus. There will be many terms with few or no index entries associated with them and a very few terms that have many index entries. This implies that the user query interface will need to guide users to the most fruitful terms, including broader or related terms, to use for their needs.

The database needed to manage these term lists and map their relationships is complex and provision will need to be made for providing a simpler interface for indexing and query programs. The prototype version of the *BioCASE* thesaurus runs on a MySQL Database and uses a Java application

programming interface (API) to handle external access. Other approaches could use database views or a reporting wrapper program.

Ongoing use of the thesaurus will develop and refine its value for relating terms and informing query building. This will create a significant resource that will need a long-term management strategy and provision made for its continued maintenance and custodianship.

INTRODUCTION

This paper discusses the factors that will influence the design and maintenance of the thesaurus required in any system that seeks to provide common access to natural science collection and observation data in Europe. The discussion is built upon the experiences of the *BioCISE* and *ENHSIN* projects and early work in the successor *BioCASE* project. The discussion has been further informed through parallel developments within the UK National Biodiversity Network. This paper concentrates on issues of design and management while details of actual proposed data structures and protocols are covered in other papers published under the *BioCASE* project¹.

150

THE ROLE OF A THESAURUS IN EFFECTIVE DATA RETRIEVAL

The continuing rise in the availability of inexpensive computing power and sophisticated software has given a growing number of organisations and individuals the opportunity both to add to the online global information resource and to access it. With cheap, fast Internet access, database owners can now reach out to both existing and new types of user and have the means to disseminate information in ways hardly thought possible even a decade ago.

Exponential growth of the Internet has gone hand-in-hand with the development of technologies, programming languages and standards that make it possible to link disparate data sources and present them through a common interface. After some 30 years of continuous effort applied to the creation of individual computerised databases, used to catalogue museum collections and capture field observations, the current challenge is how we can make use of these developments to get our information out to a wider audience.

Aside from organisational and copyright issues there are many hurdles to be overcome. These include:

- ♦ The data may be held on a variety of software applications and run on different operating systems.
- ♦ The data may be structured in different and sometimes conflicting ways.
- ♦ Similar data items (e.g. dates or geographic spatial references) may use different syntax.

¹ *BioCASE* www.biocase.org/

- ◆ Descriptive terms, taxon names, place names and other key terms may refer to different standard lists or may be idiosyncratic.
- ◆ Terms and information may be entered in different languages.

These problems have long been recognised and are part of the rationale behind the drive towards the development of national and international data standards². The development of portal software, wrapper programs and use of formats such as XML coupled with developing data exchange standards is doing much to solve the problems of linking systems that use varying software, data structures and even data syntax. Metadata standards are being developed to aid in the indexing and localisation of data sources and data standards continue to be developed to help cataloguers and recorders to ensure that their efforts are compatible with those of others.

There remains, however, the issue of retrieving the right information from this potentially vast and partly historical resource. How do you know what to ask for and if you do, how do you know that you have found all the relevant records? It all comes down to the use of words and the quality of the indexing. This is a common problem with all databases; once data are entered you cannot easily see what is in there. The data can be easily scanned for a single table with a handful of rows, but with large relational structures holding thousands or millions of records it is impossible to know directly what is in there. We are removed from the information and have to use various tools to 'fish' for it. If we know what was put in we then have some idea of what to ask for and some measure of the success of retrieval but otherwise we are working in the dark. Simple word indexes are not enough because there may be many alternative terms used, including those with a broader or narrower context, and there will inevitably be typographic errors. There is also the problem that the terms and syntax used by the cataloguer or indexer may not match those of the user – our system must meet the needs of both.

The understanding of words and their semantic relationships is not trivial but is essential if we are to address the needs of the broadest spectrum of

² Data Standards refer to the agreed organisation and content of information.

Their scope includes:

- ◆ The individual data attributes which may be organised into fields and tables to create a database. These attributes go together to form **Data Content** standards.
- ◆ The terminology used to control the content of individual attributes (e.g. taxon names or collection methods). Terminologies and structures (e.g. hierarchies) form **Data Classification** standards. Controlled terminology comes under the heading of **Vocabulary conventions**.
- ◆ The format or syntax of data in different attributes (e.g. personal names, dates and grid references. Also measurement units used) form **Syntax conventions**.
- ◆ Storage and transfer formats which describe the actual way data are represented electronically (e.g. data formats and file structures).
- ◆ **Metadata standards** for describing the content and format of individual datasets.

users. Our retrieval system must equally be able to meet the needs of specialists who know precisely what they want (e.g. the location of cultures of a specific strain of bacterium) and the child who wants to find out about dinosaurs. Our thesaurus needs to understand technical classifications and their relationships to each other and to common language.

Too many projects overlook the fact that users need the freedom to explore and the reward of discovery. The first step for users of all levels of ability is usually the question, 'I wonder what is in here?' To answer this question implies taking the language and concepts that are familiar to the enquirer and being able to produce a meaningful return, then from the familiar we can lead the interested on to the new and unfamiliar or let them follow the course of their own discoveries. This approach leads us away from the 'classical' database model, which envisages a single but refineable query that produces a definitive answer to the so-called 'berry-picking'³ approach with an evolving set of related queries used to 'explore' the resource of data that falls within the users realm of interest. 'Berry-picking' works in much the same way as one might use a web search engine to pick up clues from one search to jump to the next ('surfing'), but in this instance the user would be guided by the network of broader, narrower, equivalent and related terms provided by the thesaurus.

152

The problem of reconciling the language of specialists and non-specialists with the terms indexed from databases grows as you link databases together and more so if data are entered in different languages. This is why we need to use a special kind of thesaurus for effective data retrieval.

CONTROLLING AND UNDERSTANDING TERMINOLOGY

Most database implementations try to optimise the chances of retrieval by requiring data enterers to use key words selected from controlled terminology lists. Museum natural history and field observation databases typically use controlled classifications of taxa, minerals, habitats and place names but there are few other widely used terminology standards to cover areas such as recording and collecting methodology, specimen type, specimen preparation or even descriptors used for sites or specimens. Even where terminology is controlled, as in the use of taxonomic names, use may not be consistent and may vary through time. The inconsistency encountered in single databases rises as you try to link disparate systems into an information network.

Throughout the world there are innumerable organisations and specialists at work developing authoritative classifications and nomenclatural standards for their own specialisations. In taxonomy the *IOPI*, *ITIS* and Species 2000 projects are leading in the attempt to make available valid name lists for species worldwide. The Alexandria Digital Library, Getty Place Names

³ Bates, Marcia J. *The Design of Browsing and Berry picking Techniques for the Online Search Interface*. Online Review 13 (October 1989): 407-424.

Thesaurus and the North American National Imagery and Mapping Agency have been carrying out a similar effort for the world's place names including linking to related names and spatial 'footprints'. There are international standards (ISO series) that cover items such as country names, measurement units, date formats and others but most other standards such as habitat names tend to be national or regional in nature (e.g. the UK National Vegetation Classification and the European *EUNIS* habitat classification) and may be of restricted use. Many databases rely on 'home-made' or '*ad hoc*' term lists with no information on the source or scope of the terms and names used.

The development and stabilisation of terminology standards is of great help to those developing new databases and undertaking new documentation projects and their widespread adoption will make future data sharing easier. But even so there will never be a time when all databases are consistent in their terminology.

The key issues are:

- ♦ Term lists are rarely complete and databases have to allow for the addition of new terms. This can easily result in the addition of a plethora of closely related or overlapping terms as well as inevitable orthographic variants.
- ♦ Not all data fields needed for retrieval can be easily controlled by term lists – it is exceptionally difficult to enforce integrity rules and checks on place names⁴, person details and bibliographic references. Part of this arises from the source information being entered – do I know that Mr. Smith, P. Smith, and Peter Smith are the same person? Or that a plant from the Roman Camp and another from Leigh Woods were collected in the same place?
- ♦ Descriptive fields are rarely controlled and therefore subject to inconsistencies arising from individual interpretation of guidelines, use of terms and basic spelling errors.
- ♦ The cataloguer may be transcribing species or place names (or other data) direct from old specimen labels and may not have the time, expertise or references to check or update them.

153

If, as we have established, there are many people working on the creation and maintenance of standard classifications and term lists and that even so databases are, and will continue to be, to some extent 'individual' in their terminology, what is the purpose of creating yet another thesaurus for database networking and access projects?

The purpose of a thesaurus in this context is to enable users to effectively retrieve information across a range of disciplines, possibly in a range of

⁴ Although gazetteers of place names are widely available there are many issues involved including language, spelling and alphabet variants, political name changes and varying interpretation of the spatial extent of local place names by recorders.

languages and allowing for the ability to offer alternative search terms that match, include or approximate to those entered by the user. It is not the purpose of the thesaurus to be correct, comprehensive or authoritative, nor to act as a terminology standard – although all these things could be true for all or parts of its content. The entries in the thesaurus will be derived either from other classifications and term lists or from the indexing of partner databases. Ideally, as the data access system is used there would also be a way of monitoring user-entered queries and capturing new terms and their meaning into the thesaurus 'knowledge-base'⁵. The job of the thesaurus managers is to try to link this variety of terminology into meaningful relationships that can inform the data indexing and retrieval process.

A key function of the thesaurus will be to enable different user community views of the same data. For instance taxon specialists will expect to access data using formal taxonomic names and may be very specific in their requirements, whereas general users and non-specialists may wish to use common names (probably in their own language) and may prefer broader group terms, some of which will be informal (e.g. 'poisonous plants'). The thesaurus will, therefore, need to provide a network of related terms and concepts, ideally with weighted associations that can be explored by users starting from terminology that they understand, and which can lead them to broader, narrower and equivalent terms or related terms that might produce fruitful results. To achieve this the new thesaurus will need to bring checklists, classifications, dictionaries and other thesauri together in new relationships to power the tools used for the interpretation of metadatabase entries, database indexes and the language used by enquirers.

154

FACTORS AFFECTING THESAURUS DESIGN

PURPOSE OF THE THESAURUS

The thesaurus discussed here is a relational lexicon or faceted thesaurus, specifically designed to improve the effectiveness of data retrieval from a heterogeneous network of natural sciences databases. To achieve this it will be necessary both to supply lists of 'fruitful' search terms and to understand the scope and relationships of terms picked up by indexing programs or submitted by users (if a free term search is allowed). There is a need to cover equivalent, broader, narrower and overlapping term relationships and, where applicable, links to related terms of different types (e.g. a link from Lepidoptera to survey methods applicable to Lepidoptera). This is precisely the sort of thesaurus that will be needed by the European biological collections access service (*BioCASE*) project.

⁵ This might be important for capturing informal terms and group terms e.g. 'shore birds' or 'Mediterranean islands'. There would need to be a mechanism whereby such terms were defined and related to term lists already in the thesaurus.

A COMPLEX DATA MODEL FOR MANAGING THE THESAURUS

The thesaurus that will serve *BioCASE* will need to draw on many types of term list and have a structure that can incorporate all of their various features⁶. The principal design objective for the thesaurus management model is that it is intended as a mechanism for storing and managing many term lists and versions of lists covering all relevant disciplines (e.g. taxonomy, gazetteer, biotopes, museological terms etc.) together with the means for translating or relating from one to another. The long-term management of the thesaurus means that it will also need to store information on the source of terms, ownership of lists and any arrangements made for use or update.

The raw terms for this thesaurus will be derived from many sources and in some cases it might be important to preserve the context of the original source. For instance, the same taxon name can have different scope and relationships in different classifications or versions of classifications. This may be unimportant to the general user but of great interest to the specialist. Equally, place names, such as 'Germany' can represent quite different spatial areas at different times and in different gazetteers.

The thesaurus will need to be structured to capture many types of lists and classifications, which may have quite different approaches to sorting and expressing the relationships between terms. There may be a need to manage other information related to terms, such as abbreviations or codes, bibliographic references or other items that may assist in data retrieval or the distinction between similar terms with different applications. The core structure of the thesaurus should be capable of extension to allow for the addition of further 'added-value' data such as definitions, facts or images (or links to this information), which though not part of the current project could feature in future applications that make use of the network and the thesaurus. The structure should not preclude the possibility of others building upon the achievements of this and related projects.

Candidate models for the thesaurus structure exist. The UK NBN Taxon Dictionary⁷ was specifically designed to manage multiple checklists of taxon names related to the British Isles, including common names and versions of lists. It does not set out to be the custodian of original lists or claim any kind of taxonomic authority although it has developed into the most complete list of taxa that occur in the British Isles. For practical purposes it is also developing a 'super list' of known preferred names for UK taxa, which, though derived from other more specific works, could become a *de facto* standard in its own right.

⁶ See Appendix 1 (Term lists, indexes and thesauri) for details of different kinds of controlled terminology list structures.

⁷ Currently managed by the Natural History Museum, London. The structure of the species dictionary is briefly described at www.nbn.org.uk/projects/standard/taxdict.htm

The UK NBN Taxon Dictionary has been used as the basis for further development work under the *BioCASE* project, which has extended the NBN Data Model and established a working prototype thesaurus on a MySQL relational database at the University of Southampton⁸. The *BioCASE* prototype thesaurus is complex and includes modules for managing source, contact and bibliographic data.

A SIMPLER MODEL FOR USERS

A relational database built to meet this specification, optimised for the capture of information without loss of detail or relationships, is unlikely to be convenient for data retrieval purposes. For instance, a user query that wished to discover information about 'Roman Snails' might need to discover the formal taxon name, any synonyms and other language common names to retrieve all the likely data. This would involve some complex SQL involving iterative table joins and may not be fast. The task is increased in difficulty if the search is made to include broader terms (e.g. other *Helix* species) or narrower terms (e.g. named varieties). The structure of the thesaurus would also allow (data permitting) broadening the search to related concepts such as 'other European edible snails'.

156 The way proposed to deal with this issue is to separate the data management and data reporting structures. This can be done in a number of ways including:

- ♦ Creating pre-selected and simplified views (or copies) of the thesaurus that make query construction easier.
- ♦ Placing a wrapper program between the query and the thesaurus, which will interpret user queries into the format needed to extract terms from the thesaurus without the user needing to know the internal structure.
- ♦ Providing a full application-programming interface (API) that would enable the writing of routines to search, edit and update the thesaurus.

These approaches are currently being tested on the prototype thesaurus in the School of Biological Sciences, Southampton, in collaboration with colleagues at the University of Paris who will be testing functions for accessing the thesaurus and adding terms derived from database indexing⁹.

⁸ A paper describing the BioCASE Thesaurus Model is available at www.biocase.org/Doc/Results/WP4/D9CompleteDraftThesaurusModel.pdf and further details are on the Southampton Biocase thesaurus site www.biodiversity.soton.ac.uk/biocase/thesaurus/tim.shtml

⁹ See report for BioCASE Deliverable 9: Indexing. www.biocase.org/Doc/Results/WP5/D5IndexingModuleDescription.pdf

POPULATING THE THESAURUS

COLLECTING THESAURUS CONTENT

The *BioCASE* thesaurus will borrow from or link to terminology standards where they exist and where they are relevant to data retrieval within the *BioCASE* project. The addition of terms to the thesaurus will not be a guide to their validity only their utility. The *BioCASE* thesaurus will carry no guarantee that its included term lists are comprehensive, although it will draw wherever possible from the most accurate and comprehensive sources available¹⁰.

The proposed thesaurus will be a kind of 'knowledge-base' that knows about words and phrases applicable to natural science specimens and field observations. The scope of this knowledge is effectively unlimited but the work will essentially proceed in three phases. In the early stages the thesaurus will be populated with information (terms and relationships) derived from a wide range of existing term list sources. The next phase of growth will come from terms derived through indexing partner databases and from metadata records. The third phase will come from picking up terms entered by users and from adoption or definition of new terms to fill gaps (e.g. defining the makeup of colloquial group terms such as 'waders'). There will not be a specific end to any of these phases.

The information included in the thesaurus will be obtained in various ways. Some will be freely available in the 'public domain' and much will be derived from existing websites, publications or database applications. In a few instances lists will be assimilated directly into the thesaurus files, but it is more likely that most datasets will need to be modified to change the data structure before assimilation into the thesaurus. In a situation where a desired dictionary or data subset does not already exist in a usable form, dictionaries may need to be commissioned *de novo*. An example of the latter is the current lack of standard or extensive term lists for collection and recording methodologies and biologically related museological terms.

Fig. 1 illustrates the diversity of sources that the *BioCASE* thesaurus will derive its term lists from. It is the role of the thesaurus management team to identify and acquire sufficient term lists to provide a sound basis for indexing partner databases and to provide a framework into which new terms can be fitted. It is not the role of the thesaurus managers to validate or alter imported lists although the physical structure may be modified to suit the *BioCASE* Thesaurus Model. The thesaurus managers will have to negotiate use of some lists and arrangements for update with list owners, and all terms and data imported into the *BioCASE* thesaurus should have enough associated metadata to indicate their origins and any constraints attached to use.

¹⁰ A paper discussing the criteria used for selection of term lists for the *BioCASE* Thesaurus is available at www.biocase.org/Doc/Results/WP4/D4ThesaurusCriteria.pdf

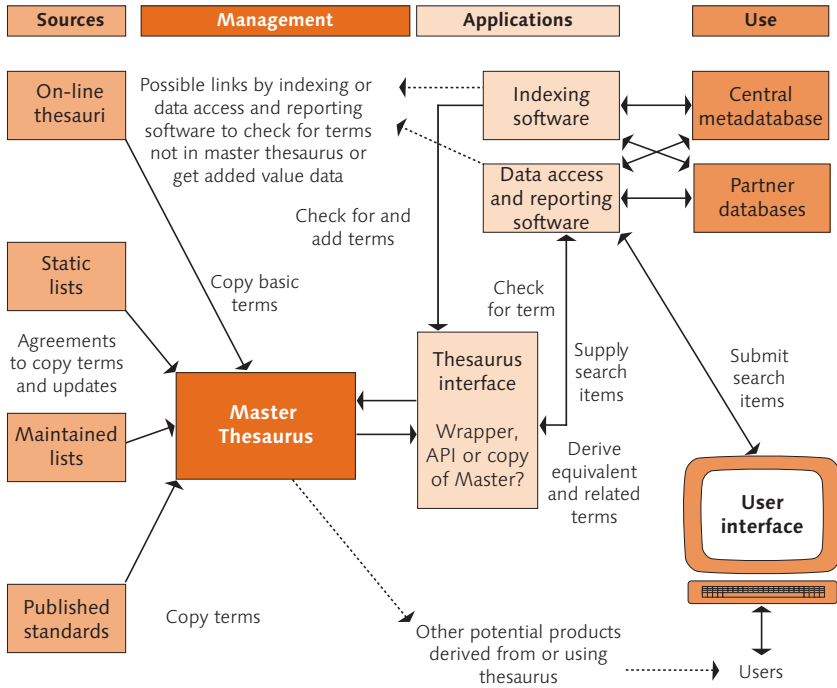


Figure 1. Schematic diagram illustrating the relationship of the master thesaurus to both sources and users of terms.

DEALING WITH THE TERM 'EXPLOSION'

The potential size of a data retrieval orientated relational lexicon such as the *BioCASE* thesaurus is huge. Both the NIMA world gazetteer and the Alexandria Digital Library Gazetteer contain over five million entries each and available world taxon lists could easily run into the millions (the Species 2000 prototype checklist released on CD contains more than 500,000 scientific and common names)¹¹. The proposed thesaurus structure allows for the recording of many linked parallel lists and versions of lists covering all aspects of natural science collections and field recording; this means that even the prototype version, created before input from indexing of partner databases will be very large. It is likely that the *BioCASE* thesaurus will contain at least seven million terms derived from other thesauri, gazetteers and checklists alone, before the addition of terms from database indexing.

Database indexing will generate many new terms. These new terms may relate to local controlled lists or could be freely entered terms. There may be many variants of existing terms due to use of varying case, plurals, gender or abbreviations as well as the inevitable typographic errors. Across the *BioCASE* area free-terms such as common names will include many language variants.

¹¹ See www.sp2000.org

Smart indexing software can help reduce the number of new terms returned from free text fields by removing trivial words and 'stemming' others for checking against the thesaurus but this can still leave many unrecognised candidate terms. It is likely that even with an extensive base thesaurus the indexing process will create an 'explosion' of terms that will need to be checked manually.

Having a very large thesaurus is not a problem, of itself; it should ensure that as many terms as possible are recognised and placed into relationships with other terms that will make for better retrieval. However, the majority of terms in a large thesaurus will have no index entries against them and if users can simply pick names or words from the full thesaurus they will be met with negative returns more often than not. This can be very frustrating and can lead to negative views about the use or quality of the service. Ideally rather than presenting enormously long 'pick lists' to the user, software should aim to guide the user to terms that will produce positive results and offer ways in which searches can be usefully broadened or perhaps narrowed when the number of returns is likely to be very high.

Broadening of searches and seeking closely related terms with positive results can be essential for maximising the effectiveness of the service for users. A simple test carried out, for this paper, on a large collection metadata-base illustrates some of the problems associated with indexing collection descriptions. The collection metadata was derived from the *FENSCORE*¹² database and consisted of 15,365 collection metadata records. A program was written to atomise the descriptive text to single words and to remove trivial words, punctuation and symbols such as brackets. The 'stop list' was manually refined until the indexing program produced a reliable list of meaningful words. All words were in English or were taxon names and no attempt was made at stemming the words to remove plurals or other variants. The resultant list contained some misspellings.

Fig. 2 shows the results of the test. The 15,365 collection metadata records yielded a total of 4,454 different words of which less than 50% appeared in the prototype *BioCASE* thesaurus (350,000 terms at that time). These 4,454 terms related to 51,086 individual index entries. All terms had at least one entry as they were derived from the *FENSCORE* data. Of these terms, 2,618 (circa 59%) had only one index entry per term while the top 25 terms (3.5%) accounted for 36% of all index entries. The top term ('fossils') accounted for 2,268 entries (4.4% of the total). The top ten most common indexing terms were fossils, plants, birds, minerals, rocks, Lepidoptera, Carboniferous, Mollusca, eggs and Coleoptera. Between them, the top ten terms accounted for around 25% of all index entries.

¹² Federation for Natural Science Collection Research – a project which started in the 1980's to document the distribution and content of natural history and geology collections in the UK. See <http://fenscore.man.ac.uk/>

Although this test is not rigorous and is related to metadata rather than unit records (which would yield many more specific taxon and place names) it does give an indication of two likely facts:

- ◆ Despite the potentially vast number of terms that could be indexed, the indexes will resolve to a relatively low number of terms with many index entries and a high number of terms with very few entries. This pattern matches well with the so-called Bradford Distribution¹³, well known to librarians in connection with the spread of articles related to a single topic across a possible range of journals (many in a few and few in many).
- ◆ The shape of the curves in Figure 2 are approaching logarithmic, which suggests that the size of the indexes would need to grow hugely to accommodate all the index-able terms that we might find but for most of the terms, queries will return very few records or no records at all. Work by other researchers¹⁴ suggests that the curve is s-shaped and will eventually plateau, but for a project the size of *BioCASE*, the plateau could be very high.

160

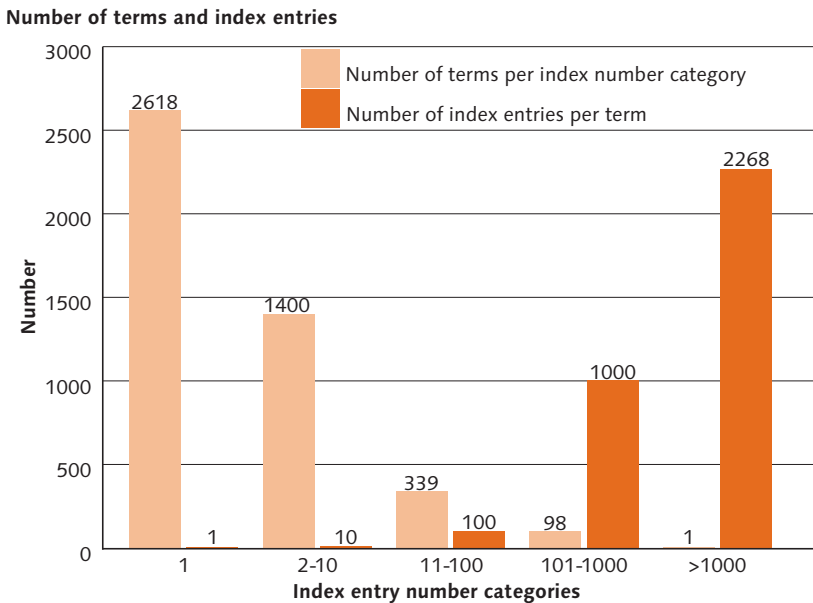


Figure 2. Results of an indexing test on a set of 15,365 collection metadata records (The UK FENSCORE Database), which yielded 51,086 index entries for 4,454 non-trivial terms. See text for details.

13 e.g. see Marcia J. Bates, 2002. *After the Dot-Bomb: Getting Web Information Retrieval Right This Time*. First Monday, volume 7, number 7 (July 2002), URL: firstmonday.org/issues/issue7_7/bates/index.html

14 e.g. see F. W. Lancaster, 1986. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, VA

It is not possible to dispense with the addition of terms for which there may be very few index records as there will be some users with needs that are this specific. What the situation does suggest, however, is the need to find ways to guide other users quickly to the most fruitful terms for their needs and also the imperative that the thesaurus can group and relate low return index terms into broader or related units that are more meaningful to users.

MANAGEMENT OF THE THESAURUS

IMPLEMENTATION OF THE THESAURUS

The thesaurus database will be very large, which will have implications for its management. File sizes may be too large for simple databases (e.g. Microsoft Access) to handle and implementation will be limited to powerful systems. The prototype *BioCASE* thesaurus has been established using a MySQL Database. Access to the thesaurus will be principally via the Internet and so provision needs to be made for both providing a user or software interface and for protecting the security of the master thesaurus database.

MANAGING THE THESAURUS CONTENT

161

A tool such as the proposed *BioCASE* thesaurus will not maintain itself. In the early stages much of the work will be involved with refining the design and populating the database with suitable term lists.

Once the main development is complete and the thesaurus is in use it might be updated automatically with new terms derived from indexing static lists and relationships will stay functional. However, added terms will need checking for validity and for linking to existing terms (e.g. as synonyms, broader terms and narrower terms). It will be impossible to manually check or relate every item so strategies and software will need to be developed to aid this process.

Dynamic lists will need monitoring for updates and new lists will be identified from time to time for inclusion. Any links to other online thesauri such as gazetteers will need monitoring to check for changed links.

LONG-TERM ISSUES

Creating a thesaurus on the scale of the one discussed in this paper also creates a responsibility and a resource requirement for its long-term upkeep and development. Upkeep implies maintaining the relationships developed with both list suppliers and ongoing access for thesaurus users. The thesaurus will be an enabling technology for the network and thought will need to be given to who will take on the responsibility for it once any initial funding projects, such as *BioCASE*, finish.

The resources and effort required to maintain and develop the thesaurus can be justified by making provision for its long-term availability to partners

and by developing collaborative ways for making its content more accurate and more widely available.

There could be a great many potential uses for such a natural sciences thesaurus and any associated 'value-added' data (especially the links within and between topic lists). Copies or extracts of the lists could be made available for new cataloguing and recording projects to increase the standardisation of future information gathering. The combination of terms and relationships that will be available in the thesaurus potentially opens up a bigger 'market' than that for any of the individually included datasets. Such information could, for instance, be of interest to search engines and directory sites for providing better location of websites.

Projects such as *ENHSIN* and *BioCASE* will demonstrate the value of networking natural science databases and the power of the thesaurus for aiding access to information that improves our understanding of species status and geographical distribution. Success will lead to greater use and refinement, which in turn will encourage new partners to engage with the project and help ensure its continued development.

OWNERSHIP AND COPYRIGHT

162

Published checklists and classifications used within the thesaurus will have copyright attached to them, likewise the thesaurus will also attract its own copyright as an original compilation, particularly if it includes researched information on origins and quality of checklists plus added-value information such as term relationships.

Ownership of each set of information (classification or term list) included in the thesaurus will need to be recorded and it is probable that a written licence to reproduce and distribute that information may need to be obtained from the original compilers of the information or authors of the classifications used. It is doubtful whether this has ever been done for the widely used classifications, which are essentially in the public domain and so the process of establishing ownership and use are as yet unclear. It is unlikely that these matters can be fully resolved quickly but should be investigated in due course. Each list or item imported into the thesaurus should be related to a 'meta-record', describing how it was obtained and any details of scope, validation or restrictions on use.

While consideration is being given in a wider context to the format of copyright and licensing agreements, it should be seen as a priority to establish the origins of term lists supplied to the thesaurus in advance of actually pursuing written permissions to use and distribute them.

APPENDIX 1: TERM LISTS, INDEXES AND THESAURI

Any organised non-repeating arrangement of words or phrases constitutes a term list. Term lists occur in different formats with different uses.

1) **SIMPLE TERM LISTS.** A simple term list is a list of words or phrases linked to a single theme. Simple term lists are used to guide data entry and provide index terms for consistent data retrieval.

2) **CLASSIFICATIONS.** Classifications link terms into hierarchical relationships where each term can have broader, narrower and equivalent terms linked to some organising concept such as evolutionary relationship, chemical composition or physical structure. Classifications may have their own non-alphabetic sort order and strictly defined rules on the creation or alteration of terms (e.g. taxonomic names).

3) **DICTIONARIES.** Dictionaries (also called lexicons) list words in a given, usually alphabetic order and may supply a variety of supporting information including meaning or description, examples of use and associated facts.

4) **KEYWORD THESAURI.** Thesauri act as guides to term use and usually classify terms into preferred and non-preferred terms ('Use' and 'Use For' relationships). Thesauri do not have to be classifications and may be organised alphabetically. They may however, link terms in hierarchical orders with broader term and narrower term categories. Thematic thesauri may include several term lists on related themes and individual terms may be related between lists (related term), it is also possible for terms to be poly-hierarchical, in that they may have more than one broader term as a direct parent.

5) **FACETED THESAURI.** A faceted thesaurus organises its collection of terms into characteristic groups called facets. Possible facets for collection indexing terms might be taxonomy, geography, and collection method. A faceted thesaurus may be thought of as consisting of multiple hierarchical taxonomies (i.e. the broader-term/narrower-term relationships within each facet), but with the possibility of relationships between terms in different facets. It is also possible for such a thesaurus to be developed as a 'semantic network' where relationships between terms can be typified. A typified relationship explains the nature of the relationship rather than using the simpler related term or broader term/narrower term of plain thesauri e.g. is a component part/has component part, or, parasitises/is parasitised by. Faceted thesauri with typified relationships are frequently referred to as ontologies. The term ontology is typically used in 'knowledge systems' to describe the definition of the inter-relationships of 'concepts' within a subject domain.

6) **INDEXES.** Indexes arrange terms derived from an indexed source in alphanumeric order and carry a reference back to the source. Many terms for the

thesaurus may be derived from indexing partner databases. Some of these terms will be derived from classified fields such as taxon name or place name while others will be derived from descriptive fields and the context may not be clear. Indexing algorithms normally atomise data into single words and may carry out further operations such as 'stemming' to reduce the number of terms by removal of tense, gender and plural versions of words. Single word indexing can lose meaningful combinations of words.