Taxonomies for business

Alan Gilchrist
CURA Consortium and Associate Consultant

0.1. Resumen

Se discute una nueva aproximación al problema de la sobrecarga de información en las corporaciones empresariales. Este enfoque se denomina "taxonomía corporativa". Se basa en las clasificaciones y tesauros tradicionales, pero usados de modo más pragmático, ya que, al constituir un sistema relativamente cerrado, puede controlarse y ajustarse mejor a las necesidades corporativas. La práctica actual de construcción y utilización de taxonomías corporativas se ilustra a través de una prospección de estudios de casos realizada por TFPL Ltd. Se realizaron entrevistas presenciales y telefónicas con veintidós grandes empresas, entre ellas Glaxo Wellcome, Microsoft y Unilever.

Palabras clave: Taxonomías corporativas. Clasificaciones. Tesauros. Ontologías corporativas. Información y documentación en la empresa. Glaxo Wellcome. Microsoft. Unilever.

0.2. Abstract

A new approach to the problem of information overload within the corporate enterprise is discussed. This approach, called the 'corporate taxonomy', is usually an extension of the traditional classification and thesaurus, but is used in a more pragmatic way and, because within a relatively closed system, can be subjected to greater control and fine tuning to meet corporate needs. Current practice in the building and use of corporate taxonomies is discussed with reference to a case study survey conducted by TFPL Ltd. Face-to-face and telephone interviews were conducted in 22 large enterprises, including Glaxo Wellcome, Microsoft and Unilever.

Keywords: Corporate taxonomies. Classifications. Thesauri. Corporate ontologies. Business information and documentation. Glaxo Wellcome. Microsoft. Unilever.

1. Introduction

Before going any further, the word "taxonomies" in the title needs qualifying; and later in the paper a tentative definition will be presented. To most people the word will mean, quite simply, classification; and will probably be most closely associated with the work of Linnaeus. However, this paper will be concerned with a relatively new phenomenon, which is beginning to be known alternatively as the "corporate taxonomy", and specifically with the results of an investigative research project designed to survey current practice in the building and application of these tools. The Project was carried out by TFPL Ltd., a company serving the library and information sectors, with offices in London and New York (www.tfpl.com). In the past few years TFPL has been building a strong presence in the areas of Information Management and Knowledge Management, providing a specialised recruitment service, training and consultancy in many aspects of IM/KM. This interest in IM/KM naturally led to work with intranets and portals; and in turn this identified a growing area of concern with, and activity in, the corporate taxonomy as a tool for combating the problem of information overload at the desktop.

2. Information overload

A recent study from the University of California (Varian, s.d.) on the amount of information currently being generated has, for the first time, it is claimed, used terabytes as a common standard of measurement. (A terabyte equals one million megabytes). The authors have produced some staggering statistics:

- The directly accessible "surface" Web consists of about 2.5 billion documents and is growing at a rate of 7.3 million pages per day. Counting the "surface" Web with the "deep" Web of connected databases, intranet sites and dynamic pages, there are about 550 billion documents, and 95% is publicly accessible.
- A white collar worker receives about 40 e-mail messages daily at the office.
- Original documents created by office workers represent nearly 90% of all original paper documents, while 56% of magnetic storage is in single-user desktop systems.

The TFPL study certainly confirmed that the typical office worker is now subjected to a flood of documents, either passively or proactively. For example, there are some 2.2 million documents on Microsoft's intranet with, of course, the ability to access the Internet as an 'extra' information resource. The next wave in increased quantities of information is likely to be generated by the advent of electronic commerce, notably B2B. One estimate (Norton, 2000) suggests that UK B2B revenues are likely to rise to £28 billion in 2002, some 4% of the UK GDP

(£28 billion is roughly 7,500 billion pesetas). Much of the associated information will flow through extranets.

There are many technological and administrative problems to be solved on the Internet and on intranets, but senior management is only just beginning to be aware of the serious nature of the problems at the human level. Information overload caused partially, ironically, by the intranets that were designed to bring information to the desktop is causing frustration, and in some cases, clinical stress. This, in turn, leads to a drop in both efficiency and effectiveness as workers hunt for the appropriate information in the electronic jungle. To make matters worse, it is now being recognized that there is also a real problem of information literacy, which is only recently beginning to be seriously addressed (Webber, 2000). Most workers, while being computer literate have a very poor idea of how to conduct effective searches, nor how to store, manipulate and use the information retrieved. For example, it has been found that some 70% to 80% of electronic searches are conducted using one keyword! The corporate taxonomy is seen as being part of the answer to the twin problems of information overload and information literacy.

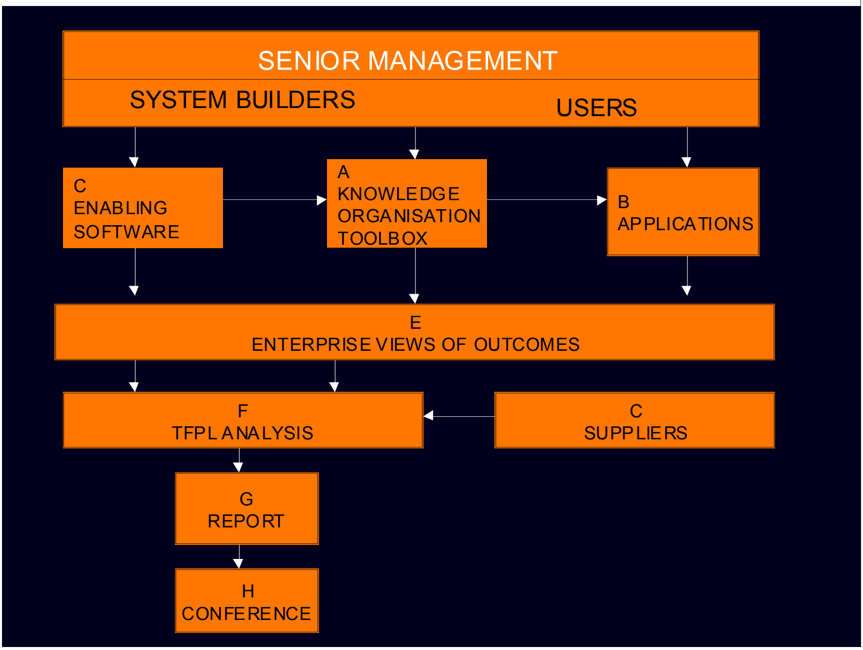


Figure 1

3.The research project

It was decided early on that the project would not offer a definition of the word taxonomy for two reasons; first, it was thought that the word in its new context was not yet sufficiently current and second, that the project would attract a wider range of views if it started from a statement of the general problem of information overload, as was outlined in the section above. Consequently, a number of large enterprises was approached, who it was known had installed intranets, and some of whom were known to be active, or interested, in building taxonomies. In effect, the opening question was "We are interested in finding out whether you are experiencing the problem of information overload at the desktop, and what you are doing to combat the problem". Referring to Figure 1, the project really started at Box B - Applications, in order to find out what each enterprise had in its toolbox at Box A, and what enabling software was being used (Box C) and from what suppliers (second Box C). It is worth stressing here that the project did not set out to evaluate software or suppliers though, inevitably, much was learnt about this aspect. Finally, the project team took the view that in a project of such a short duration, there was no time to get involved in direct investigation and evaluation, so that the entire survey was conducted through the systems builders and their managers, wherever available. The project was conducted throughout August and September 2000, and was based on face-to-face interviews in the core case studies and by telephone in the supporting case studies. The six core studies were of the British Broadcasting Corporation, Glaxo Wellcome, the Ministry of Defence, together with its database manager, the Defence Evaluation Research Agency/Defence Research Information Centre (MoD/DERA-DRIC), Microsoft, PriceWaterhouseCoopers, and Unilever. There were 16 supporting case studies, some of which supplied so much information by phone and email as to be classed as core. The draft Report was presented to a Conference in London in October, and the final Report appeared at the end of November 2000 (www.tfpl.com).

4. Evolution of retrieval languages

Because the project started with no definition of the word taxonomy, it was able to collect a range of approaches to the common problem of information support at the desktop; and consequently it was possible to discern an evolutionary upward trend in a hierarchy of complexity, as organisations were forced to deploy stronger tactics. Figure 2 provides a broad perspective on this, expressed in simple language for a lay reader. The points made are that:

- Classification and traditional taxonomy are, essentially, the same and can be regarded as either act or product, supported by some mode of language-based access, and largely based on pre-coordination.

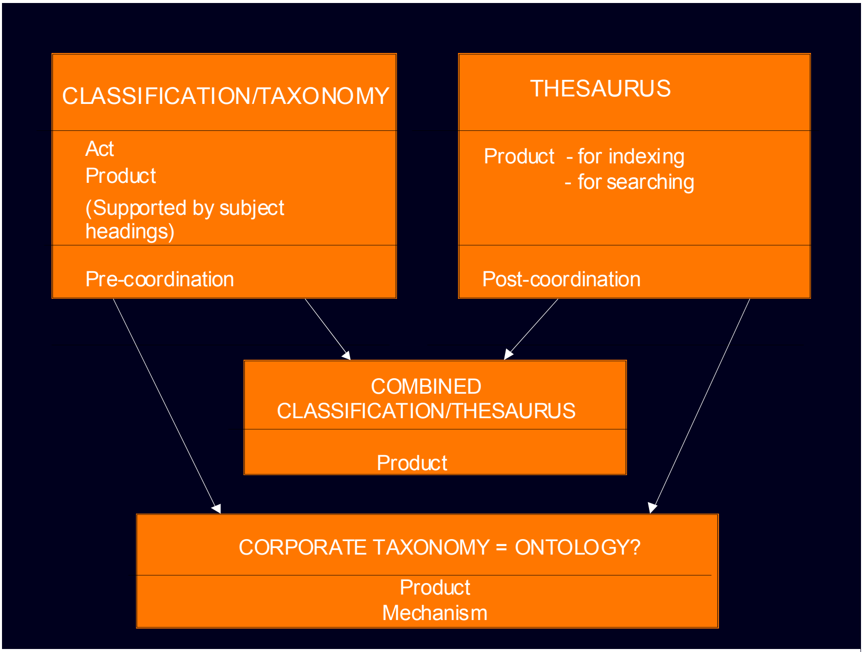


Figure 2

- With the advent of mechanical methods, notably the computer, the thesaurus became prominent, used for indexing/searching in post-coordinate mode.
- For a brief period, the classification and thesaurus approaches merged, most famously with the Thesaurofacet (Aitchison, 1969), but this was largely overtaken by the introduction of full text searching and direct access to the end user, made possible by distributed networking. It would seem that the "honeymoon" period of end user access to full text is now (possibly temporarily) in decline for the reasons outlined in the section on information overload above.
- Now, the corporate taxonomy is proposed as a tool, which might be seen as a tool or, as will be implied later in this paper, as a mechanism; and one which is likely to employ both classification and thesaurus building.
- Though the project discovered several "tools" which could justifiably be termed as taxonomies, the most ambitious share certain characteristics with ontologies. For example, in a paper discussing the relevance of ontologies to information scientists, Vickery (1997), quotes Alexander et

al. "…a methodology, called ontological analysis…which consists of an analysis tool and its principles of use, that result in a formal specification of the knowledge elements in a task domain". A corporate enterprise is a task domain which employs a range of functional languages, and if the employees of that organisation are to talk freely to each other (particularly in an age of globalisation), then new tools must be constructed.

5. The case studies

5.1. MoD/DERA-DRIC

This case study illustrated the evolutionary point made in the previous section. DRIC has the responsibility of maintaining databases for DERA and the MoD, of which the largest is the database of research reports. The original collection, now totalling some 300,000 documents has been indexed using the original U.S. Department of Defense Thesaurus, compiled in 1967, and supported by the COSATI Classification, both of which have been continuously updated by DRIC. Following a reorganisation within MoD and DERA, a further 850,000 research reports were added, which had been indexed by a different scheme, obliging DRIC to effect a rough reconciliation between the two. (As will be seen later, this problem of mergers between organisations and schemes is one of the prime triggers for taxonomy creation). At a later date the function within MoD responsible for administration of internal and commissioned research realised that it needed a greatly improved scheme of documentary control, and asked DRIC to adapt its thesaurus for the indexing of the register of research projects. At the same time, it realised that it needed a classification scheme to support policy making and resource allocation, and so it set about devising a taxonomy. This is a three level hierarchy, having just three main classes, as can be seen in Figure 2. The first class is concerned mainly with "materials", the second with "equipment" and the third with whole "assemblies". Managers of research projects are obliged, each year, to code their research projects with an A Code, and B and C Codes, if these are known; or a B Code, if that is the starting point of the research. Each Class is accompanied by copious scope notes, and the taxonomy has been proven to be highly successful. In fact, it has now been adopted by all UK defence industry organisations in the National Defence Industries Technology Strategy.

The final twist in this case study is that MoD are now planning and beginning to implement a system called the Knowledge Store, through which all employees of the MoD and DERA will be able to access any database mounted on their intranet. Discussions are now under way regarding the compilation of an organisation-wide taxonomy to support this facility, and it will be interesting to see how the DRIC thesaurus and the MoD taxonomy will fit into this new scheme.

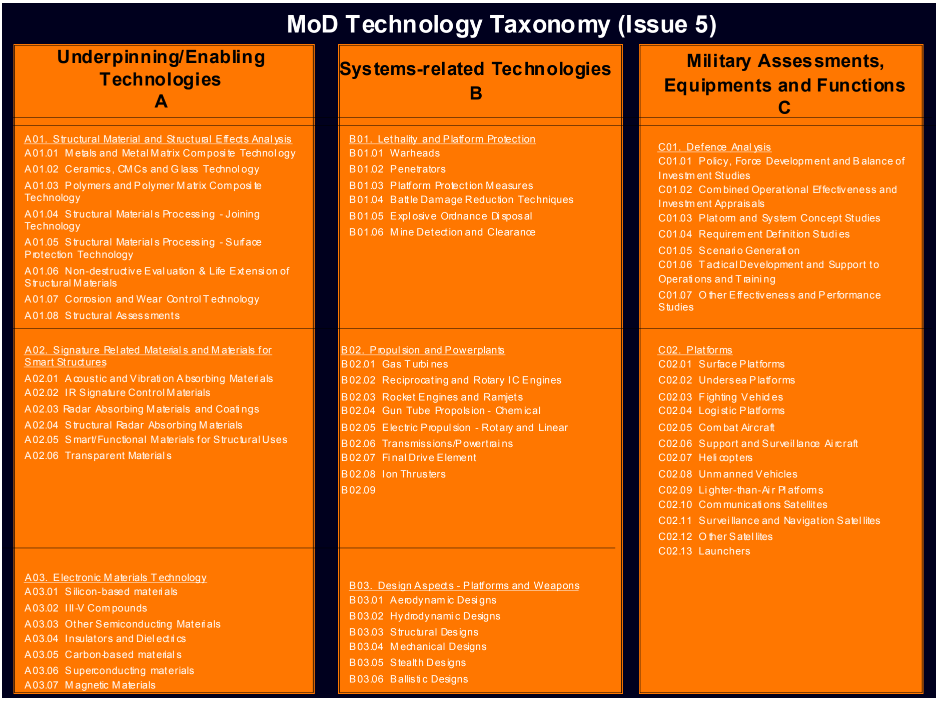
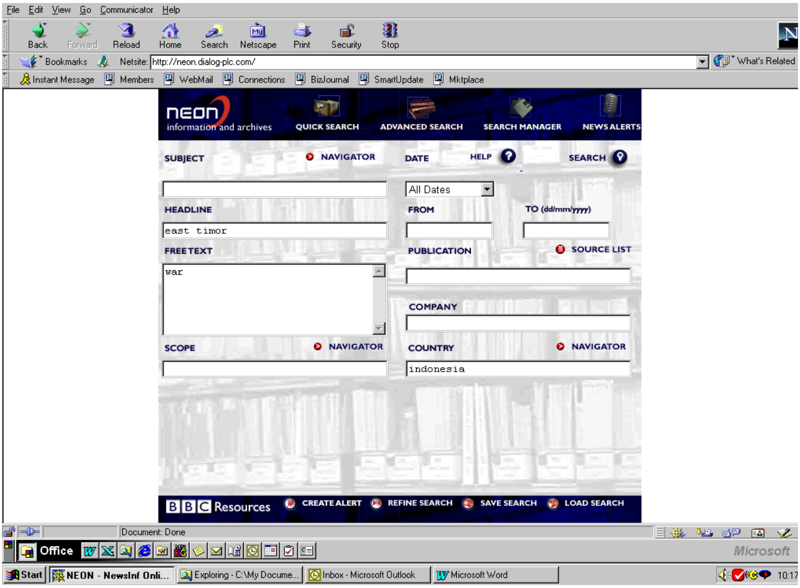


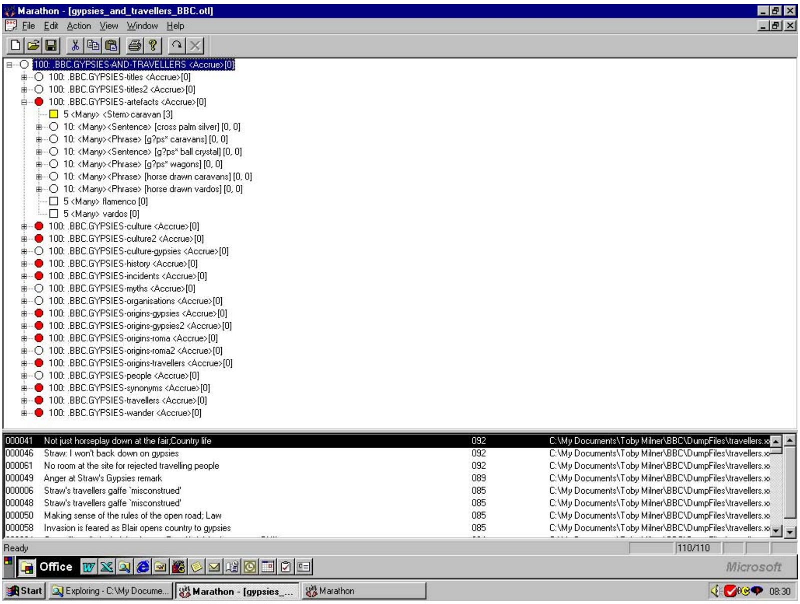
Figure 3

5.2. BBC

Following major reorganisation in the BBC some few years ago, a new unit was established, which brought together a wide range of information resources, previously scattered and with autonomous management. This created an enormous single resource, called Information & Archives, containing news cuttings, TV and Radio broadcasting video and sound recordings, music (both recorded and sheet music), bibliographic material and internal records associated with programme making. All of these repositories contain material dating back many years, including for example video and sound recordings from the early days of broadcasting in these media. Access to this single collection poses enormous problems in terms of size and media format, and two major problems have been identified as:

- how to modify the legacy systems and indexes;
- how to refine the indexing process, i.e how to reduce its labour intensive nature.





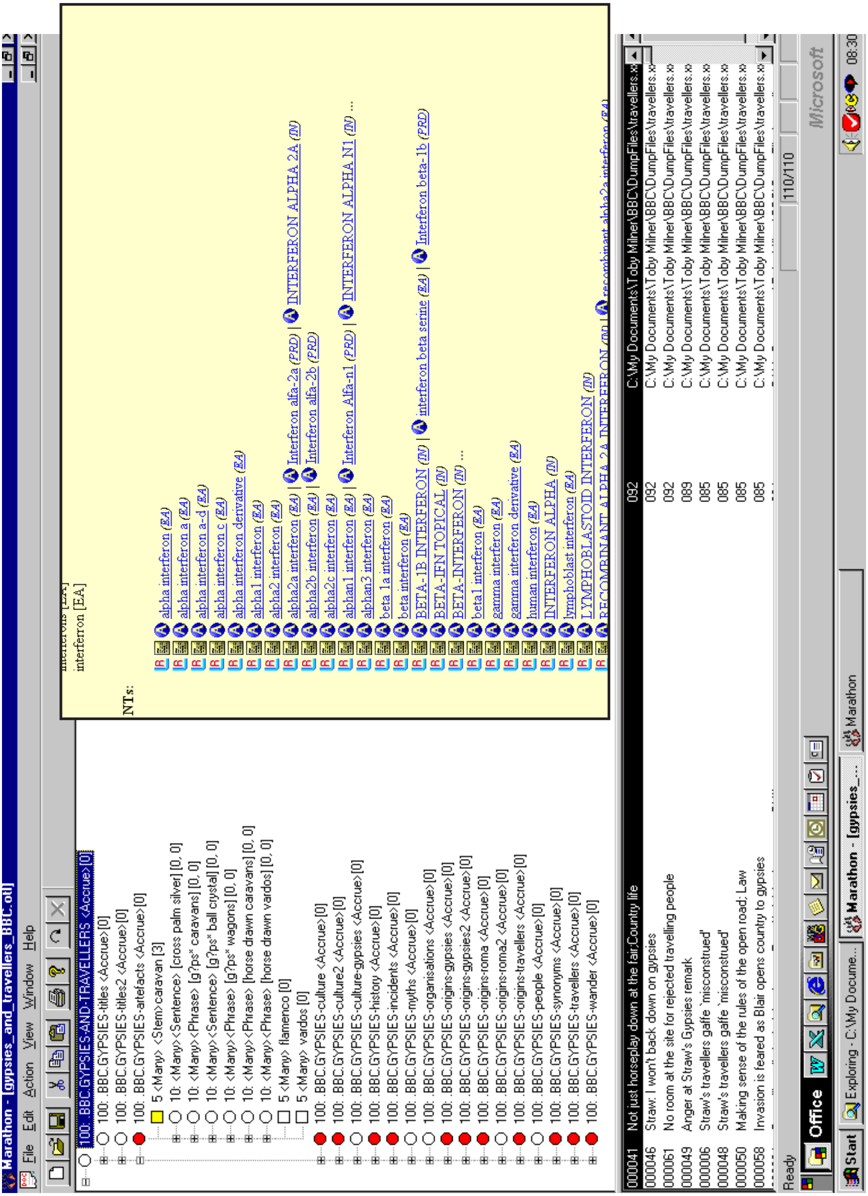
Figures 4 and 5

It is interesting to note that the indexing process is to be refined, not abandoned, as the BBC have conducted tests to show that human indexing is still necessary in their environment.

Tackling all the different media types in one single operation was clearly impracticable, and as it was early recognized that the area which offered the greatest strategic advantage in starting a development programme was that of newspaper cuttings, it was decided to work in this area to demonstrate two separate benefits: a significant savings in costs accompanied by addressing the information needs of a high-profile user population.

The newspaper cuttings collection, which had been accruing over a period of some 70 years, had reached by the mid 1990s the colossal total of 20 million cuttings, growing at the rate of 2,500 every day. This labour-intensive exercise involved the scanning of the major daily and Sunday newspapers, as well as selected serious weeklies and specialist magazines, marking up and cutting out indexed items, (from duplicate copies where necessary) which were then stored in bundles of cuttings, each bundle being filed under the first subject in a string of headings and sub-headings. Moving to an electronic environment, which now meets most of the need of news-based programme makers, the costs of provision have been reduced to about a fifth of the previous spend. User satisfaction is reflected in the fact that 5,200 passwords were issued in the first year, and use is increasing.

NEON (standing for News Information Online) is a portal using the AltaVista search engine, enhanced by a taxonomy created with the software package Infosort. This is a rule-based package, which serves to expand a search statement according to the rules that have been built up in the rule-base by the knowledge engineers (a combination of Infosort experts and BBC staff, familiar with the terminology of the collection). Figures 3 and 4 illustrate search screens wherein various search keys may be entered, including a “Controlled” subject term (in the example RACE AND POLICE). The rule-base will contain information about this term within the context of the collection, such as synonyms, closely or remotely related terms (controlled by weights), word form (plural/singular; stemming etc.). Figure 5 shows a fragment from the rule-base concerning the keyword “Gypsies”. Infosort operates according to probabilistic indexing algorithms, calculating on the basis of the statistical occurrence and co-occurrence of words and phrases. Early experiments in automated rule-base building appear promising.



Figures 6 and 7

5.3. Glaxo Wellcome

The Glaxo Wellcome taxonomy, now a massive compilation, started in a relatively conventional way. A Wellcome chemist had been asked to merge two thesauri, when Wellcome merged with Glaxo, and a further five, or so, thesauri were offered for consideration. The chemist saw this as an opportunity to create something more powerful than a merged thesaurus, though this is one aspect of the tool that resulted. The implication behind a merged thesaurus is that it becomes the ultimate authority file for indexing and searching, catering on the way for specialist terminology; but Glaxo Wellcome is a huge and global business, and this brand of centralisation is likely to be unacceptable. The chemist, turned system builder, set about creating his megathesaurus as a knowledge map of the repositories and their accompanying thesauri offered within Glaxo Wellcome. Thus the megathesaurus became a correlation of the terminology used across the company, allowing a user to enter via a term "preferred to him or her", and to browse through associated terms and then on and into selected databases. This huge undertaking comprises:

- 53,470 concepts (the fundamental terms "preferred by the taxonomy")
- 201,749 lexical equivalents (including synonyms and near synonyms)
- 443,511 assertions (relationships, both hierarchical and associative; and where many of the latter are defined; e.g. Medicine X – RT (acting on) – Disease Y).

Figures 6 and 7 show an excerpt from the entry for the Medicine Interferon, containing Top Terms, Broader Terms, and Narrower Terms; followed by bracketed references to the associated databases. This would be powerful enough, but for example, for each medicine it is also possible to look up its chemical formulation, and the legal and packaging regulations associated with that product for each country in which it is marketed.

In developing this knowledge map/knowledge base, Glaxo Wellcome worked with a software house called Cycorp. This company has been working for some years in the field of ontology creation, and is known for its "common knowledge" ontology called the CYC Ontology. Put simplistically, Cycorp developed a sophisticated Thesaurus Manager for Glaxo Wellcome, but one that is rule-based. The claim is made that any new thesaurus can be automatically assimilated into the "megathesaurus" through this rule-base. Glaxo Wellcome and Cycorp have jointly filed a patent for this software.

5.4. Unilever

Many special libraries adopt a “just-in-case” approach to collection management. Unilever attempted a “just-in-time” approach and use an unusual technique in bringing together the research scientists (users) and the information scientists charged with looking after their information needs. The technique used was the “mind map” advocated famously by Tony Buzan (1995). Led by a moderator, a small group of research scientists and information scientists sat round a table on which there was a large blank sheet of paper. Starting with a major topic (Figure 8 shows the final result of a mind map starting with the topic “Cardiovascular health”) the group brainstormed a map of associated topics in which Unilever was scientifically and commercially interested. In fact, of course, many pieces of paper were used, and the first topic took 4 or 5 half day meetings before consensus was reached, though the teams were able to speed up when more confidence and experience were gained. The first results were used in a public relations exercise; but then the serious work began. The information scientists took “clusters” from each of the mind maps (Figure 9 shows the “Cholesterol” cluster from the topic Cardiovascular health, which might serve as an example) and used the terms in the cluster to prepare highly detailed and elaborated search statements,

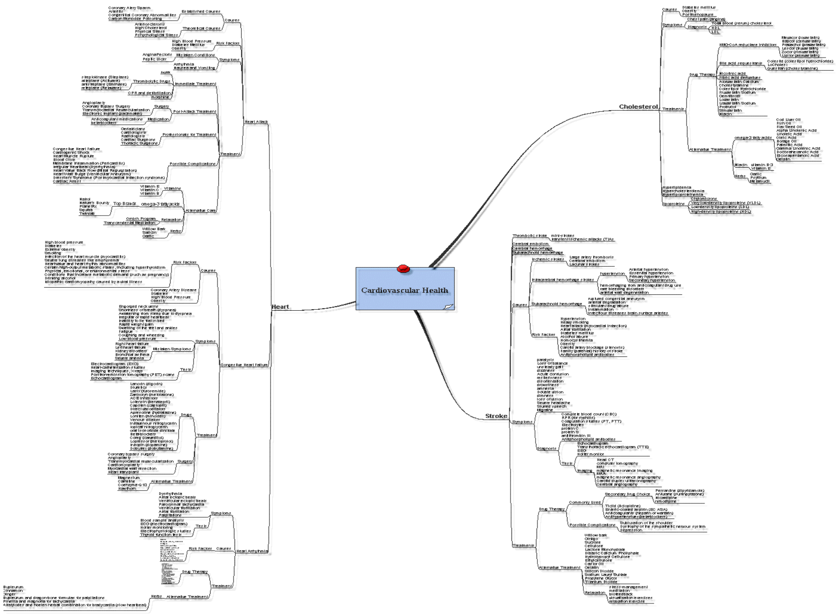
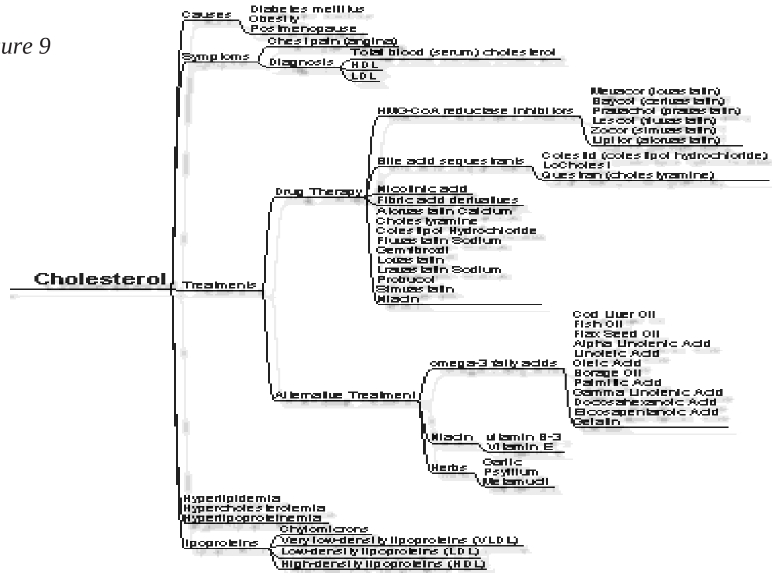


Figure 9



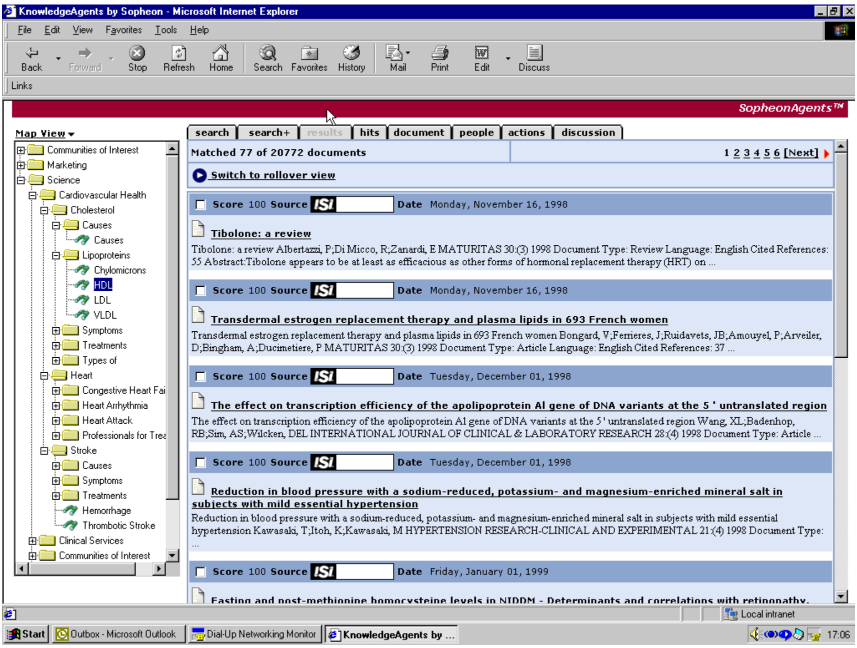


Figure 10

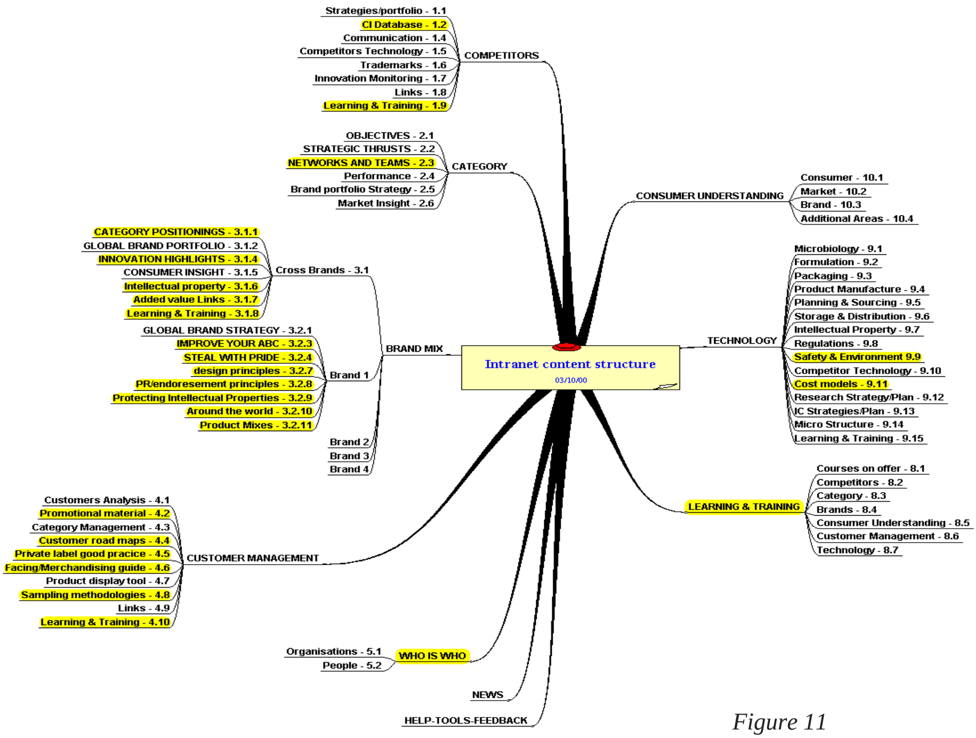


Figure 11

each one tailored to a particular database, e.g. Medline, Web of Science, Dialog. The results were then downloaded, normalised with regards format, and stored in a Unilever database. Next, search screens were devised for the research scientists to gain access to this database, and the mind maps were incorporated as knowledge maps to aid navigation. In fact, as can be seen from Figure 10, the map was reproduced in file manager format, after the users had become familiar with the radial format. This Figure shows a results screen following a search on the keyword HDL. Users have the facility to over-ride or add to the controlled keywords using other controlled terms or free text search. They also have the facility to annotate retrieved items, and to find out who else in Unilever has read the retrieved items, and then to be able to contact them direct to discuss their opinions.

The idea of mind-mapping spread from the Research Centre in the South Midlands of the UK to the Marketing function in the North-East. After early trials it was discovered that the approach had to be adapted. Whereas the Research function was essentially building a single database, the problem in Marketing was that there was already a number of overlapping and duplicating databases, caused by the fact that Business Regions were maintaining Product databases, and Product Groups were maintaining their databases with Regional information. Rather than finishing up with a topic based mind map, the Marketing function ended with an "Intranet content structure", and this is shown in Figure 11.

5.5. United States Post Office

The area of visualisation is likely to attract increasing interest, and it is significant that many software suppliers in the text retrieval and portal areas refer to knowledge mapping. One supplier, who makes a feature of the word "taxonomy" in its promotional literature is the Semio Corporation, used by the USPO.

The USPO has been an independent non-governmental agency for the last 30 years, and employs some 800,000 staff. Many of these in the marketing function are now reaching retirement age, and senior management realised that a crack programme in information management was timely.

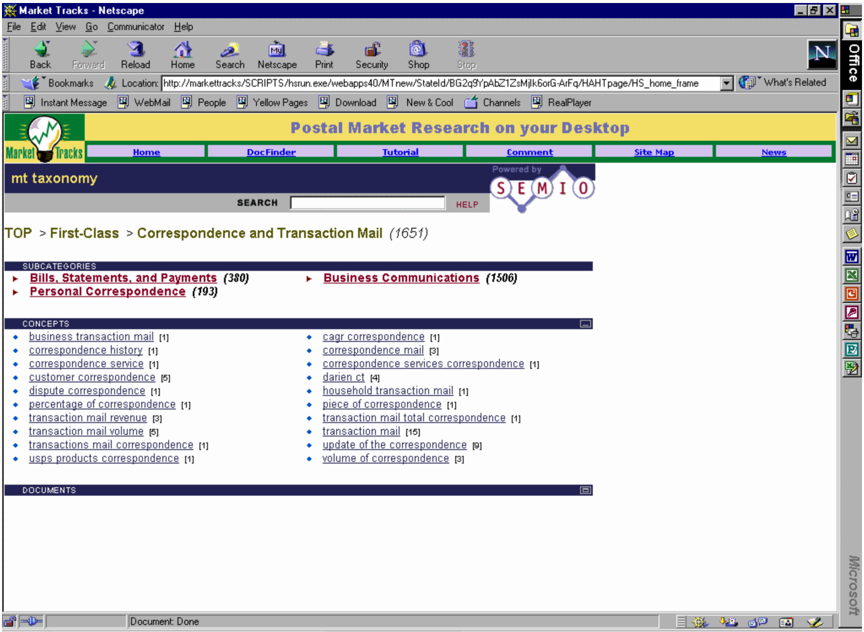
Starting with collections of business glossaries, seeded with some particular knowledge culled from staff, a taxonomy was automatically generated by Semio from the analysis of a repository of market research reports and similar material. This taxonomy can be displayed in a number of ways. Figure 12 shows a traditional menu approach, such as may be found on many contemporary public portals, while Figure 13 shows a "Semio map" automatically generated from the terminology bank.

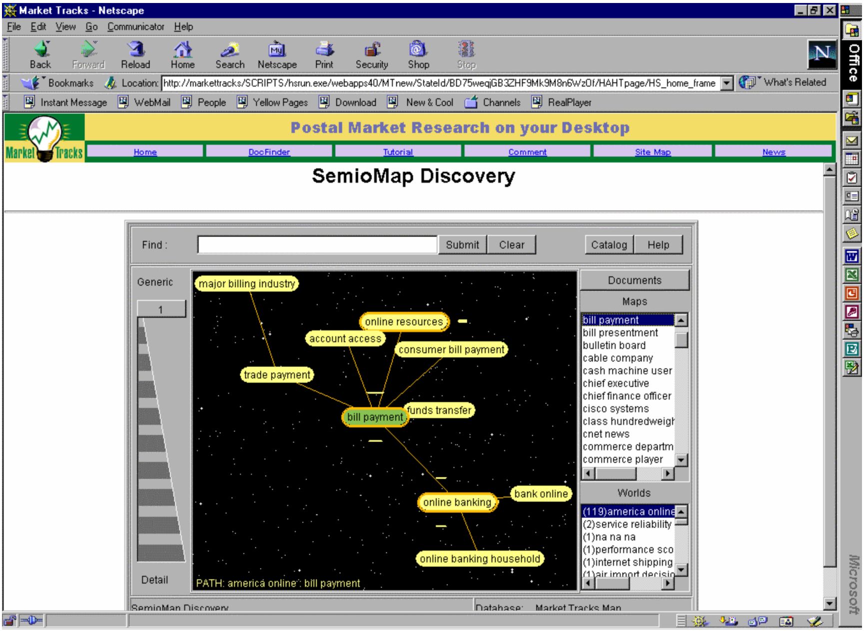
5.6. Montgomery Watson

Both of the last two case studies laid some stress on the people factor. In the case of Unilever, and as part of its Knowledge Management Programme, there was an explicit attempt to connect people to people through the bibliographic database, while the USPO example highlighted the attempt to capture the knowledge of retiring staff. In the case of Montgomery Watson, a main driver behind the compilation of a taxonomy was to compile an index of expertise. Montgomery Watson is an environmental engineering company employing some 4,000 specialists in over 30 countries. Figure 14 shows a hits page listing both primary contacts and technical experts, with one selected and highlighted for the topic of "air quality".

5.7. Microsoft

Microsoft had a similar end in view as Glaxo Wellcome, but took a more evolutionary top-down approach. There was the same recognition that a multiplicity of databases was available on the MS Intranet (containing some 2.2 mil-





Figures 12 and 13

lion documents); and that efficient and effective access was becoming increasingly problematical. As with Glaxo Wellcome it was necessary to promote collaborative effort between a number of working groups, involving portal owners and content providers.

A start was made by tagging document surrogates for high-use items identified by a combination of data modelling and user feedback. From this a complete mechanism was evolved consisting of a taxonomy covering (in principle) all material within the Microsoft firewall, and a “Central Metadata Tag Set Registry”. The architecture for this mechanism is shown in Figure 15; wherein can be seen in the middle row the metatag registry, comprising a core tag set and specialist sets for certain areas. The taxonomy resides in the bottom row, where it can be seen that there is a core vocabulary repository, and again, accompanied by non-core repositories – all feeding into the metatag registries. This mechanism is then made available to all portal owners and content providers for a variety of applications which are summarised in the top row, and include an authority file for standardised tagging, support for portal navigation through such devices as “drill-down” menus, and support for search engines. With this evolutionary

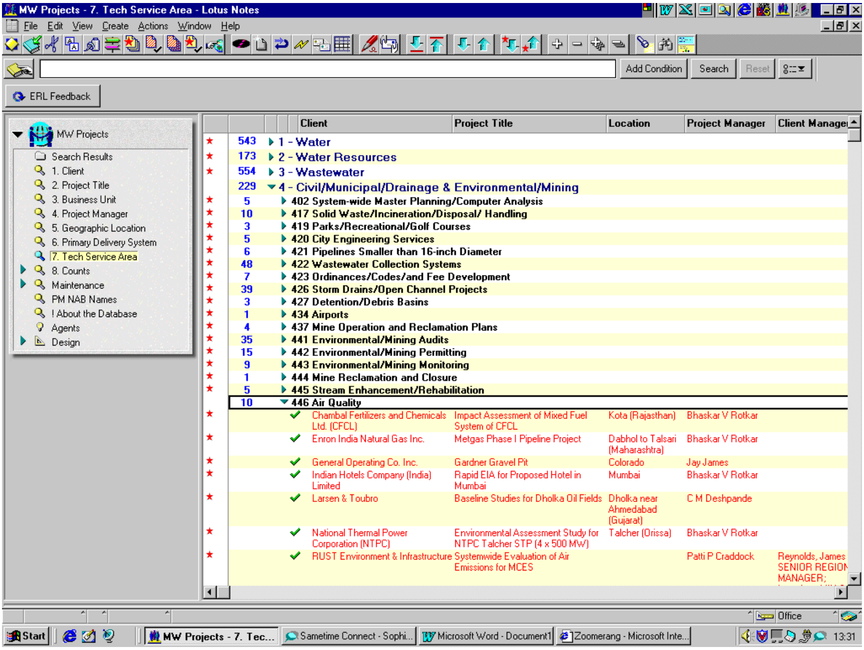


Figure 14

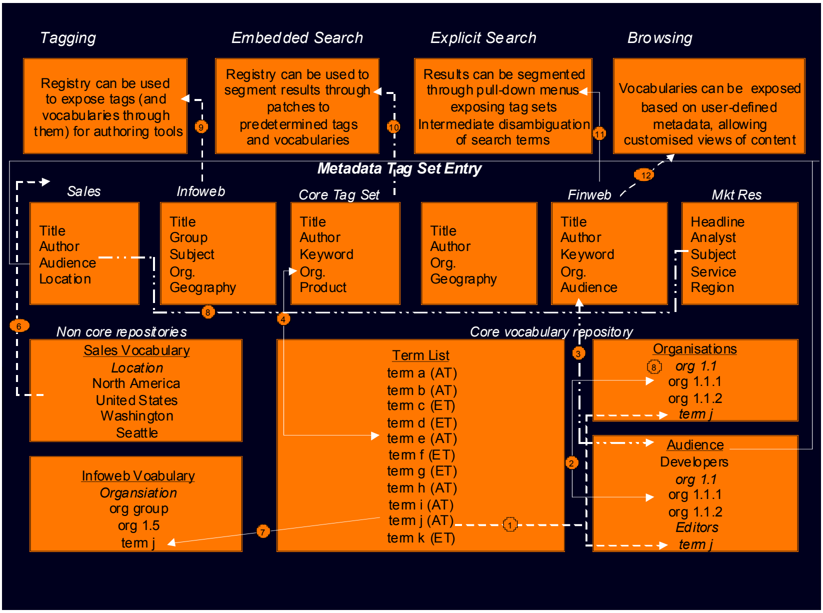


Figure 15

approach, there is likely to be a slow convergence of systems as collaboration increases and designers appreciate the value of a central facility which offers compatibility, rather than imposes standards.

Microsoft has attempted some evaluation of this system, while acknowledging both the importance and the difficulty of arriving at objective metrics. However, user statistics and feedback have been extremely encouraging, and the knowledge architect stresses the importance of keeping in tune with user needs and expectations.

6. Conclusions

There are many ways in which enterprises are attempting to tackle the problems of information overload and individual user problems with information processing. It is, however, possible to discern some pattern and perhaps a trend towards the broad paradigm offered by Microsoft and Glaxo Wellcome. One major difference between public and corporate portals is that the former operate commercial databases using published classifications and thesauri; while the corporate enterprise, acting as it does in a relatively closed environment, can afford to be more pragmatic and can exert a greater degree of control and fine tuning to meet corporate needs. On this basis the study concluded with a tentative definition of a corporate taxonomy as follows:

A corporate taxonomy aspires to be:

- a correlation of different functional, regional and, possibly, national languages used by the enterprise …
- to support a mechanism for navigating, and gaining access to, the intellectual capital of the enterprise …
- by providing such tools as portal navigation aids, authority for tagging documents and other information objects, support for search engines and knowledge maps …
- and, possibly, a knowledge base in its own right.

Subsequent work at TFPL is re-examining the data to determine whether more light can be shed on the results by looking at the "taxonomies" from the separate points of view of their structure and their function.

The most important conclusions seemed to be that:

- Businesses are now expending significant human, financial and technical effort to resource and develop taxonomic approaches that will support exploitation of their disparate information and knowledge resources well into the future.
- The majority of enterprises surveyed were looking for greater business effectiveness through improved information access; or efficiencies in information retrieval, which free time for more productive use. Intranets and enterprise-wide information portals are likely to fail if the issue of information content structure is not addressed.
- The corporate intranet has provided a common technological platform, and XML has the potential to provide a common data exchange platform. The challenge for individual enterprises now is to devise and deliver common information platforms.
- Few organisations had produced business cases for the taxonomy at the detailed level, but the great majority had backing from senior management to tackle the problem. This is a critical factor for success.
- At the heart of the taxonomy debate is the need to achieve a balance between the talent of the taxonomy designer, the cost to the system to implement the taxonomy and the familiarity of the users both with the system and the structure of the information itself.

- Most of the enterprises felt that over-reliance on software solutions was dangerous; and were, in consequence, prepared to invest labour-intensive time for human intervention in building, applying and maintaining taxonomies.

7. References

Varian, Hal ; Lyman, Peter (s.d.). How much information?. URL: <http://www.sims.berkeley.edu/how-much-info/index.html>.

Norton, J. (2000). Strategies for e-business transformation. London :Institute of Directors and KPMG E-business ; Director Publications Ltd., 2000.

Webber, Sheila ; Johnston, Bill (2000). Conceptions of information literacy: new perspectives and implications. // Journal of Information Science. 26 : 1 (2000) 3-19.

Aitchison, Jean ; Gomersall, Alan ; Ireland, R. (1969). Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects.Whetstone, Leicester, (England) : English Electric Co., 1969.

Vickery, B.C. (1995). Ontologies. // Journal of Information Science. 23 : 4 (1997) 277-286.

Buzan, Tony (1995). The Mind Map book. London : BBC, 1995.