The Business Benefits of Taxonomy

A SchemaLogic White Paper
October 2005

Judi Vernau
Director, Information Architecture
Metataxis

SchemaLogic Inc.• 620 Kirkland Way, Suite 100 • Kirkland, WA 98033 • Phone 425.885.9695 • www.schemalogic.com

INTRODUCTION

There was a time when taxonomy meant something to zoologists and botanists, but very little to anyone else. In 2001 Gartner predicted that by 2005 at least 70% of the Global 500 will deploy taxonomies in order to improve the organization of their internal and external documents[1]. This year the Chartered Institute of Library and Information Professionals set up the Taxonomy Network and 71 people attended the inaugural meeting in London, Synapse Corporation set up the Taxonomy Warehouse (www.taxonomywarehouse.com) listing over 200 publicly available taxonomies and thesauri, the UK Office of the e-Envoy identified 24 departmental taxonomies in government, and the Ark Group ran its sixth successful taxonomy conference in three years.

The power of taxonomies has been promoted by the growth in appreciation of the importance of knowledge management, and also by the realisation by many organizations of how the classification of their content can bring additional potential for revenue generation. The latter is easier to make the business case for, the former is still hard to provide a solid return on investment statement for. So, given that so many organizations are willing to make that investment, with more joining their ranks every week, how have they justified the expense?

This White Paper reviews the issues around information retrieval and the reasons why taxonomy is becoming a popular way of dealing with some of them. It looks at how the evangelists within public and private organizations have made the case for developing a taxonomy, and what benefits they expect to reap or have already received. Finally we look at approaches and technologies which will potentially help to minimise investment and so increase revenue.

Finding the needle in the haystack

> **"Our ability to create information has substantially outpaced our ability to retrieve relevant information."**[2]

Web technology makes it easy for us all to be publishers now, and while the readers of this White Paper may not necessarily publish their own content on the internet, there is a very strong likelihood that they will be making at least some of their documents available via a corporate network, intranet or extranet. A worker can put her latest report on the company network so that technically it is available to anyone to read, but will the potential reader be able to find it, and having found it once, will they be able to find it again?

What you can't find is clearly of no use to you. Similarly, if you don't know what you've got, how can you make the most of it? There are few hard facts regarding the loss to business arising from this inability to locate the right information at the right time, but in 1999 IDC estimated that US Fortune 500 companies would lose $12 billion in 2000 due to an inability to locate knowledge resources.[3] IDC also cites a communications firm which estimates that "by improving search and

retrieval systems for just the firm's 4,000 engineers, the investment would be recovered within a month and would contribute to a $2 million monthly productivity gain thereafter." Jakob Nielsen has estimated that poor classification of knowledge assets costs a 10,000 user organization $10 million annually[2].

Many commentators focus on the time wasted in searching. The consensus seems to be that 25-35% of knowledge workers' time is spent in looking for information, with less than 50% success[4].

However it is not helpful to think of losses only in terms of time wasted – this is clearly important, but it is not really possible to know or measure accurately how productively someone would use their time if they were not searching. Another way of viewing the issue is to consider what the organization is losing by not having vital information easily available. This can manifest itself in various ways:

- longer time to market as staff spend time recreating work that already exists
- lost opportunities as useful research is overlooked
- customer frustration at not being able to find information (or because staff can't find it for them)
- inability to exploit existing information fully
- difficulties with compliance with legislation on freedom of information

All of these issues point to the need to classify content in order to make it readily findable.

Why search is not enough

Three or four years ago many businesses were using search engines to enable them to solve problems of information access and retrieval. A search engine would crawl over unstructured data and return results based on text matching, word clusters, and other semantic techniques. While these methods do indeed play a valuable role as an option for retrieving information, they have significant limitations:

- Keyword search captures only 33% of relevant information (Source: Chris Wilkie, BBC Information and Archives, Sept 2002)
- "Most of the complaints we get are due to the way users search – they use the wrong keywords." (Source: Must search stink?, Forrester, 2000)
- 40% of search failures come from customers and information providers using different terms.

Search engines assume that users know what they are looking for and can use very specific keywords to be able to retrieve it. But users of information are often not sure what documents they require until they see them: they only know the general topic they are interested in. In a Yahoo! market research project, 75% of respondents preferred browsing to searching as it allowed them to view all the documents related to a particular topic. While such research may face accusations of bias, the enduring popularity with end users of the Yahoo! directory must be testament to the benefits of browsing.

An Alta Vista study showed that 80% of user could not or would not build a proper Boolean search, and 87% used less than three words. A BT survey also showed that almost half of search engine users never use more than a single term in a search. This means that search terms will return a much greater number of hits, but that no attention is paid to the fact that a single word search (or even a compound term) may have multiple meanings (for example, 'depression' may relate to geology, history or meteorology, and even French Connection refers to at least two different things). Documents containing synonyms, alternative spellings or the same concept expressed in different language will usually be missed altogether (although some search engines claim to be addressing some of these issues).

It is clear that keyword search will always be an important method of locating information, particularly when very specific and unambiguous terms are used, but in TSO's view, taxonomy supports an approach to information management which is much more consistent and comprehensive.

How does taxonomy help?

Taxonomy has been described as "a systematic classification of a conceptual space"[6], in that it seeks to encompass and provide labels for all the significant concepts within a particular domain, rather in the way that a traditional classification scheme does. A taxonomy provides a hierarchical structure for these concepts, from the broadest (for example, The Arts) to the narrowest (for example, Trip Hop), allowing users to understand the context of each label or term as they navigate through. The taxonomic structure can be applied to topics, organizations, places, or any other categories of concept which relate to each other hierarchically.

The kind of taxonomy (or taxonomies) required by an organization will be determined by its information strategy and the main drivers behind it. For public bodies the Freedom of Information Act 2002 and the e-Government Interoperability Framework are significant drivers for taxonomic development as part of a wider metadata framework, but for all organizations taxonomy can support the following activities:

- searching
- re-purposing of content
- unifying language across organizations
- future-proofing knowledge held in the business

Searching

Searching via a taxonomy may involve navigating the hierarchy, or could take place by searching for a specific term within the hierarchy (the latter will be more successful if it is supported by a thesaurus which provides additional access points via synonyms and acronyms).

In either case it is very important that the taxonomy should group topics in a way that reflects the expectations of the users, and that it should use their terminology. However, the user does not need to know the exact term for every concept, because the taxonomic structure will provide groups of related terms within the broader known context. For example, a jazz fan will easily find his way to

The arts > performing arts > music > jazz

but in looking further down the hierarchy he might find types of jazz he didn't know existed:

Jazz > avant garde
bebop
big band
cool
fusion
hard bop
mainstream
progressive
third stream
trad.

The hierarchical structure also solves the problem of polysemy, or multiple meanings, of words and compounds, because the context is immediately clear. If a user searches the taxonomy, the result might be not a hit list containing documents relating to all and every meaning of a word or phrase, but a set of potential contexts so that the user can select the relevant one, for example:

Fusion **The arts > peforming arts > music > jazz > fusion**
Science > physics > nuclear physics
Computing > application servers > Cold Fusion

BENEFITS IN ACTION

The Freedom of Information Act (FOI) 2000 gives the public a statutory right of access to government information unless there is a clear and acceptable reason why that information may not be published. This will entitle any person to be told upon written request whether a department holds particular information, and if it is held, to have that information communicated to them within twenty working days.

Michael Warner of the UK Ministry of Defence has been working on a taxonomy which is currently being implemented across the whole of the organization, and which is expected to greatly faciltate the search for requested information. "Although it is not known how many requests will be forthcoming, it will undoubtedly impact on MOD business resources." Michael says. "The advent of FOI places a new emphasis on the

importance of good record keeping, being able to identify what is or is not held and where it is located. Using controlled subject terms from the taxonomy … will be of considerable benefit for the subsequent retrieval of the information requested in the timescale required." Michael reports that the benefits of the taxonomy are already being felt by users looking for information in the daily course of their work. "We had excellent buy-in from management all the way along, but the biggest selling point for [continuing to develop] the taxonomy now is the fact that they can see that it's working].

The controlled terms are providing a common framework of concepts (and relations between these concepts) to structure MOD's business language. Our objective is not only to provide the list of terms for use in writing and information seeking, but also to create maps between concepts to connect staff with the right knowledge at the right time. It constitutes an essential tool for managing intellectual capital and connecting MOD staff with knowledge."

Re-purposing

It is clearly sound economic sense to re-use wherever appropriate content that we have already paid to create. The trick is to have labelled that content in such a way that

- items can be found again
- items can be recombined into useful information sets.

Ensuring that items can be found again means not only giving people efficient access to information, but also obviates reinventing the wheel on a regular basis. In a 1998 study Kit Sims Tayler reported that knowledge workers spend more time unwittingly recreating existing knowledge than in creating new knowledge.

"NASA, like all federal agencies, needs to make the best use of workers' time. When an engineer or scientist finds and reuses content, the return on investment (ROI) for the time and effort to originally produce the material increases. The cycle of creation and reuse directly impacts the Agency's operating costs. It also pushes the pace of development forward at a greater rate as teams build on previous work instead of "reinventing the wheel" over and over again."[7]

The recombining of the different items can be made possible by classifying each one according to a predefined vocabulary. This means that all items relating to jazz, for example, could be gathered together for a research project or publication.

If the vocabulary is derived from a taxonomy, the potential for broader and narrower combinations is increased, as all content about music can be grouped, or just those items relating to jazz, or just those items relating to the trumpeter Miles Davis.

In the presentation of the material to the end user, the taxonomy also has a useful part to play, as the hierarchical structure can be used as a means of arranging the content and allowing the user to navigate through it.

BENEFITS IN ACTION

National Health Service Estates publishes a series called Health Building Notes (HBNs), which provides regulations and guidance for the construction of buildings for the health service. Currently the HBNs are only available in print, as a result of which a great deal of information is repeated in each publication. More importantly it can be difficult for users to find all the information they need relating to a particular aspect of a building unless there is a specific HBN dealing with it. As a result of this, NHSE is proposing to develop a set of taxonomies to describe the different aspects of health buildings, and the appropriate terms from the taxonomies will be assigned to each paragraph within the series. By classifying each paragraph in this way, the information becomes a resource which can be re-presented in a number of ways, with each paragraph needing to be created only once. In future, users will be able to request all information relating to a specific clinical discipline (cancer care, orthopaedics, etc), space type (office, waiting room, ward, etc), user type (in-patient, disabled, etc) or topic (lighting, ventilation, etc), thus providing a much more responsive and efficient service to the user.

Unifying language

A key issue for organizations that are spread over several locations, are made up of a number of self-contained units, or are the result of a merger between two or more organizations is the variation among its constituents in their methods of handling information and in their use of language to describe them. One department may choose to arrange its documents by business function, another by topic. The US team may refer to Human Resources, the UK team to Personnel. The key to successful retrieval of the information is to find a way to make all the labels consistent, so that searchers can be confident that they have found all the available material. The development of a taxonomy (or if necessary, a suite of taxonomies to cover different facets of the information) is the way to ensure this level of consistency so that everyone can use the same language.

BENEFITS IN ACTION

In October 2000 the UK Government published the e-Government Interoperability Framework (e-GIF), which mandated the adoption of internet and world wide web standards for all government systems. Part of the e-GIF is the e-Government Metadata Standard (latest version 2.0, May 6 2003), which among other metadata fields includes the requirement to use at least one term from the Government Category List (GCL).

Local taxonomies are also encouraged, and can be mapped to the GCL so that the people assigning the subject terms only need to use these local terms. This means that departments and agencies can describe their local information as precisely as required, but that there will always be a common language across government.

Future proofing

Once an organization has begun to store and share its information assets according to a well-structured taxonomy, the impact of losing staff is lessened, since at least the explicit knowledge can be found by others.

Of course, the taxonomy must continue to grow and change if the organization is going to keep up with current and future developments and be able to respond effectively to changing user requirements.

BENEFITS IN ACTION

INUK (InvestUK) operate in the competitive area of national inward investment, attracting investors to the UK, and particularly parading the benefits of the UK compared to its European neighbours. Responses to the 700 enquiries each year from abroad cover transport, education, economic conditions etc and need to be prompt, well-prepared, interesting and focused on the specific opportunites.

Jointly owned by DTI and FCO, and working with Trade Partners in a proposed British Trade International portal, INUK have been working with a set of investment factors and a set of industry sectors to describe their content, but these vocabularies are not used consistently or to a sufficient level of detail. They are now proposing to introduce a new editorial resource for creating and re-using investment material, and a new information labeling discipline is required that will classify new and existing material by investment factor and provide for rapid compilation of responses. A more comprehensive taxonomy will be introduced that will change as markets shift and new differentiating factors emerge. It will also be capable of incorporating vocabularies from new markets. This taxonomy will be built into their internal processes and also be used in commissioning and updating third party material, and hopefully be adopted in the Regional offices as well closely related parts of DTI and FCO.

Winning the case for taxonomy

How do you persuade the Board, and in particular the Finance Director, that the organization needs a taxonomy? If there is already a Knowledge Management strategy in place, the argument is clearly easier to win, as the company already understands the value of information, at least at some level. But when you are competing for budget, how can you make a compelling case for something that doesn't immediately seem to be revenue generating or cost cutting?

Return on investment is usually calculated by comparing the investment cost, which is measurable, against predictions for increased revenue. In the case for taxonomy, unless it is being used to support a new saleable product, it is very difficult to give a meaningful estimate of financial benefit. But taxonomy is not alone in this: how does the Board measure the benefit of a network upgrade or of subscribing to a major news service? The main selling point of these kinds of investment is that they can increase productivity and competitiveness. Is taxonomy any different, or is it just that it seems like the latest trendy buzzword that will go out of fashion in due course and is therefore not worth spending money on? Or is it that people associate taxonomies with library classification, and feel that something so 'old-fashioned' can't really be worth serious investment?

One way of combating both these views is to bypass the word taxonomy altogether. Some taxonomy champions report much more success when they refer just to 'common vocabularies', or use commonly understood examples such as Yahoo!, which is more usually described as a directory structure.

Organizations often seem more willing to introduce new technology than to invest in something which is seen as so labour intensive as a taxonomy. In fact, as we will suggest later in this paper, development costs need not run into six figures, and there are ways of keeping expenditure down. But it worth remembering that giving people the latest email software, for example, is fine, but if the company provides no way of locating colleagues' email addresses, it is of no benefit.

So in making the business case it is important to tailor your argument and the terms it is couched in to suit the audience. In particular it is useful wherever possible to find out what special information issues your stakeholders are facing. In the following paragraphs, we look at the different types of benefit that can be listed as part of the argument, and wherever possible we underpin them with statements from organizations who have found some way of measuring results. Following that, we look at the types and levels of potential costs.

Increased efficiency

As stated earlier, research suggests that knowledge workers spend 25-35% of their time searching for information with only 50% success. Microsoft has developed a taxonomy for all its internally created information and for material licensed from third parties. Mike Crandall, Microsoft's Knowledge Architect Manager, as reported in a NewsEdge article[7], has stated that the company has growing evidence of the value of the taxonomy based project: "Even at this early stage, they have seen a 62% reduction in the number of clicks, an average of 16 seconds saved per task, and an 11% increase in task success rate."

If the average knowledge worker earns, for example, £45,000 (including overheads) and spends 30% of their time searching (at a cost therefore of £13,500), and if the number of clicks can be taken as a measure of productivity, this would save the company over £8,000 a year per person.

Vicki Hooper of Unipart has estimated that the company has saved £2m in time and productivity through the introduction of a taxonomy. Jonathan Engels of Reuters has said that that company expects to save £40m by developing a better taxonomy. The United States Postal Service reports that five years ago a research project into the effect of changing postal rates took 350 hours to complete. After the introduction of a taxonomy to classify all inhouse and third party data, it took one person just two hours to find the appropriate information.

Another approach to evaluating increased efficiency is to consider how much more effectively information can be managed if it has been properly classified. Knowing that you have leads to at least two potential savings: ensuring that the same material is not created twice (or more!), and being able to rationalise the amount of documents stored. We have not been able to find any reports of actual time and money saved by not reinventing the wheel, but the benefits of rationalising storage space are easy to calculate. The Department of Work and Pensions recently carried out an asset cleansing activity which reduced the number of documents by approximately one third. This kind of deduplication, apart from making search and retrieval much easier, can also lead to the decommissioning of servers, which could easily save £x [CHECK] per server.

Re-purposing

When assets are properly classified, that classification can be used as a strong foundation for creating new groupings of material for dissemination, as evidenced by the National Health Service Estates example given above. The advantages can be calculated in straightforward terms of cost/benefit analysis: for a relatively small publishing project it might cost, for example, £20K to create the taxonomy and £10K to apply it to the content set, plus £70K for the rest of the product development, but the revenues from the new product should be at least £50K per annum, so the breakeven point is within two to three years.

This only works in the commercial publishing environment of course. But many other organizations will want to exploit their assets more fully, particularly government departments who are charged with making information more easily accessible. The public use very different vocabularies to civil servants, and taxonomies on government web sites now reflect that. Reassessment of internal documents under the Freedom of information Act is also resulting in the need to reclassify assets to make it easier to answer the kinds of requests that the public are likely to make.

Competitiveness

The United States Postal Service started to develop a taxonomy three years ago, and used ExcaliburSemio (now ConveraEntrieva) to implement it. John Gregory, Marketing Specialist at USPS, says "Having software that isolates any occurrence of a concept means we can give clients a definitive answer on what we know and how we know it. This can be done in minutes, instead of a full work day, and we can have confidence that the search was exhaustive."

Risk reduction

Risks around information management, as we discussed above, not finding what you need when you need it resulting in lost opportunities, finding the wrong thing, and being unable to find or present information to end users. The Foreign and Commonwealth Office has a history of sharing knowledge via email, but has very little centralised storage of information. FCO employees currently suffer from an inability to find 'a single version of the truth'. Difficulties with multiple versions are very common in the public sector, and the Freedom of Information Act in particular is encouraging these organizations to tackle these problems. Making sure that the public can find the information thaya re entitled to is a major driver for implementing well-designed classification and retrieval technologies.

Calculating the cost

What content set will the taxonomy be applied to? Is it a specific group of material, like our NHSE example above? Or a public website? Or an intranet? The scale of the development will differ, but the types of cost to be taken into account are the same: there is the cost of building the case in the first place, and then – assuming agreement to proceed - costs for gathering the vocabulary and agreeing the structure, probably costs for hardware and software, for creating and maintaining the taxonomy, perhaps for migrating existing content, and for training both those who will assign taxonomic terms and those who will use them to search.

Daniel Rasmus, VP of Giga Information Group says that a typical taxonomy implementation usually costs about $100K[8] and Forrester estimate $75K[9]. As a guideline, the Unites States Postal Service taxonomy, which covers 20,000 documents, took 30 days to build. A taxonomy developed by the UK Ministry of Defence to cover the MOD, Armed Forces and Defence Agencies and currently comprising just over 1,000 terms took three months of liaising with each department to understand and include the different vocabularies, using an expert to facilitate and guide the effort.

Although it is possible to build a taxonomy using something as simple as a spreadsheet, most medium to large developments will benefit from using specially designed software such as SchemaLogic. These applications add to the cost, but generally make the task of taxonomy development and maintenance much simpler to manage by offering authorised users the ability to create new terms and relationships between terms, to link terms to documents, to link multiple taxonomies together and to map to external taxonomies, and even to manage and link taxonomies in different languages.

Ways of increasing the return on investment

There are several ways of approaching taxonomy development which can improve results in the appropriate context:

- Start with a smaller, but clearly defined body of content, for example a set of documents which are easier to classify, or which are particularly important. A basic taxonomy could be compiled based on these documents, which could then be refined over time.

> "According to Forrester's [Don] DePalma, organizations have a greater chance of success if they begin with a problem that is manageable, measurable and has a well-defined scope. 'Large-scale efforts that are launched as a Big Bang kind of thing rarely succeed,' he said. 'We often suggest that organizations pick a customer-facing problem, such as cutting customer-service calls in half, because you can easily quantify that.'"[10]

- Get expert advice: finding a skilled specialist who can guide the effort may be an expense but it is one which is very likely to save money in the long run. It may also be useful to have a neutral voice to act as facilitator between the different people and departments involved.

> "The process of categorizing data need not be either expensive or overly complex. In recent correspondence on the email list xml-dev, Carol Ellerbeck, a taxonomy expert with Harvard Business School's Baker Library and formerly of Lycos, made this very point. Responding to a writer who suggested that one needed to be 'king of the world' and have 'an unlimited budget' to create effective taxonomies, Ellerbeck wrote, 'If you "were king of the world"…you would not need "an unlimited budget"...just a modest one, to have experts build your taxonomy/domain vocabularies. I say this as a taxonomist who has been in the vocabulary trenches with electronic information for years. Automation is wonderful (and I would say, even essential), but start with not just humans (albeit smart humans), start with humans who have some expertise, and you will accomplish your goal faster, with fewer people, more efficiently, and have a more solid foundation to build on." [11]

- Build on existing work: do you have your own vocabularies or glossaries? Are there existing taxonomies which could be licenced?

- Use a proven software vendor who offers pre-loaded taxonomies: SchemaLogic now offers a variety of taxonomies and thesauri which are already incorporated into its taxonomy software.

- Are there additional business opportunities which can be exploited as a result of developing the taxonomy? (There may be other products and services that can be derived from new groupings of material, as discussed above. Factiva have taken this one step further by building on their own success story to offer taxonomy consultancy and development services to other organizations.)

- Use automatic document classification: this is unlikely to be appropriate for smaller document sets, but for applying the taxonomy terms to larger bodies of material, automatic classification tools are becoming more and more widely used and with increasing rates of success. Some system vendors such as Interwoven also claim to be able to compile the taxonomy automatically, but TSO remains sceptical.

"Subject structures are far too important an area to be left to a software package with no knowledge of an organisation."

Dr Nic Holt, a Fujitsu Fellow, speaking at the Advanced Knowledge Technologies Town Meeting, 1 April 2003

If you need to use multiple taxonomies, use industry standards: authors and users are more likely to be comfortable with the vocabulary if it is taken from a known and used standard, and the terms should be easier to map to other organisational taxonomies if required.

References

1 Linden, Alexander, *Innovative approaches for improving information supply,* Gartner, Sept 4 2001.

2 *Taxonomy and content classification: market milestone report,* Delphi Group, 2002

3 Feldman, Susan, and Chris Sherman, *The high cost of not finding information: an IDC White Paper,* IDC, July 2001.

4 For example: "Workers spend approximately 25-35% of their time searching for the information they need, rather than working on strategic projects and business opportunities." Andrew Warzecha, META Group, July 31 200, and "Knowledge workers spend 35% of their productive time searching for information online, while 40% of the corporate users report that they cannot find the information they need to do their jobs." Working Council of CIOs, *Business Wire*, Feb 27 2001.

5 Jim Nesbit of Semio, quoted in di Maio, Paola, *The key to successful searches,* FT IT, Nov 7 2001.

6 Dutra, Jayne, and Joseph Bausch, *Enabling Knowledge Discovery: Taxonomy Development for NASA,* NASA Technical White Paper, Jan 7 2003.

7 Bryar, J.V., Taxonomies: the value of organized business knowledge, NewsEdge Corporation, 2001

8 Cited in Pack, Thomas, Taxonomy's role in content management, *e-Content Magazine,* March 2002.

9 Sonderegger, Paul, Taxonomies, ontologies make search mopre accurate, *Techstrategy*, Feb 27 2003

10 Roberts-Witt, Sarah, Practical Taxonomies: hard-won wisdom for creating a workable knowledge classification scheme, *Knowledge Management,* January 1999

11 Trippe, Bill, Taxonomies and Topic Maps: Categorization Steps Forward, *e-Content Magazine*, Aug 2001