Building a corporate taxonomy: benefits and challenges

Eric Woods: EXW@ovum.com

February 2004

Expert advice

Building a corporate taxonomy: benefits and challenges

Taxonomies are a fundamental part of any modern information architecture. Any organisation that needs to make significant volumes of information available in an efficient and consistent way to its customers, partners or employees needs to understand the value of a serious approach to taxonomy design and management.

However, many organisations are unfamiliar with taxonomy development and management. At its simplest, a taxonomy is a hierarchical organisation of categories used for classification purposes. Such a simple definition hides the many challenges to be faced in building and maintaining an effective and usable taxonomy for your organisation.

This report explains why taxonomies are a key issue for many organisations, and looks at the benefits they bring and the challenges to be faced in developing and maintaining a corporate taxonomy. It also examines the role of categorisation tools in taxonomy design and maintenance, and looks at future technology trends.

The value of taxonomies

According to the writer Jorge Luis Borges a 'certain Chinese encyclopaedia' classified animals as:

belonging to the Emperor; embalmed; trained; sloppy; sirens; fabulous; stray dogs; included in this classification; trembling like crazy; innumerable; drawn with a very fine camelhair brush; et cetera; having just broken the vase; from a distance look like flies.

The wondrous inconsistencies that this brings to mind are a useful reminder to anyone trying to implement a corporate taxonomy that the world does not fit easily into neatly labelled boxes. That said, it is increasingly important for those in charge of content management, enterprise search, portal or e-commerce projects to have an understanding of why taxonomies matter and how they can be used to improve information retrieval and navigation.

In some areas, such as botany or medical research, taxonomies have long been important tools for organising information. Today, many more organisations are looking to build taxonomies as part of their information management strategies. Taxonomies are an important tool in balancing the contradictory forces of information overload and the need for instant access to the right information.

Why taxonomies are important today

The recognition that we need to organise information if we are to make sense of the world can be traced back to Aristotle. Each advance in the scale of human knowledge has presented new challenges in classification and new responses to those challenges: for example, Linnaeus's system for categorising the natural world in the 18th century, the creation of the Dewey Decimal System for library classification in the 19th century and 20th century medical and scientific taxonomies. The rise of information technology, particularly the Internet and the Web, presents a further series of challenges that are extending the requirement for taxonomies into a far wider sphere.

There are three key drivers for the current level of interest in taxonomies.

Information overload

The latest Berkeley University study into information growth estimates that 5 exabytes of recorded information were created worldwide in 2002 (equivalent to 800Mb for each person on the planet). If access to these volumes of information is to be a benefit rather than a burden then order and control become prerequisites. Information management techniques must be improved if we are to gain more control over these information flows, and taxonomies should be a key part of this. (For more on the Berkeley study see www.sims.berkeley.edu/research/projects/how-much-info-2003.)

The rise of the Web

The very structure of the Web (in an Internet, extranet or intranet context) offers new opportunities for information organisation. The ability to provide universally accessible, hyperlinked, multimedia content presents unique challenges in terms of information classification. The growing awareness of the value of taxonomies is closely associated with the rapid development of knowledge about effective website design, online usability and the importance of the overall information architecture.

The growing use of unstructured information management technologies

In response to, and as part of, the evolution of the Web, most large and medium-sized businesses have invested in content management, search and portal technologies. They are now looking at how they can increase the benefits from these technologies and provide a consistent information infrastructure that can be shared across different applications.

Taxonomies are not part of the usual technology hype

More than two centuries after the death of Carl Linnaeus (the 'father of taxonomies'), it is hard to consider taxonomies a fad. However, there are inevitable concerns about how relevant taxonomies are to the needs of most organisations. Such concerns are inevitably associated with scepticism over vendor hype for new tools and technologies.

Our view is that taxonomies are a fundamental part of modern information architectures. Any organisation that needs to make significant volumes of information available in an efficient and consistent way to its customers, partners or employees, needs to understand the value of a serious approach to taxonomy management.

In many cases, organisations are learning about taxonomies the hard way through the iterative development of their Internet and intranet sites. This is an expensive and risky approach to the management of core organisational assets. Investing in a taxonomy development programme is more efficient and provides important benefits.

The benefits of developing a taxonomy

Improved quality

A taxonomy provides a basis for accurate and consistent subject metadata. As part of a general investment in the quality of an organisation's information, a taxonomy can help improve the efficiency (and reduce the cost) of application integration, website design and knowledge management initiatives.

Easier navigation, more efficient search

A good taxonomy should make it easier for users to navigate through large volumes of information. The taxonomy should also improve the results from search engines, which can exploit both the improved metadata (to increase the accuracy of a search) and the taxonomy structure (to narrow searches and organise results).

Together these improvements can reduce the time that employees or customers spend looking for information and increase the likelihood that they will find what they are looking for. A well-considered taxonomy also increases the opportunities for the serendipitous discovery of information (for example, information on similar projects or additional works by the same artist on a music site).

Improved information sharing

A common taxonomy provides a shared language for different parts of an organisation. For example, product research and development (R&D) and marketing should find it easier to share information across departmental divisions.

It can also reduce the amount of time spent on duplication and reinvention, by making the existing intellectual capital resources more visible and accessible. This will become particularly important as information management tools and techniques are expanded into new areas such as regulatory compliance. Organisations concerned about meeting new regulations on corporate governance, data privacy or freedom of information will benefit from a consistent approach to information organisation, even if they are not using the taxonomy directly.

A better user experience

More and more customers judge a company on the quality of the website and in particular the ease with which they can find the information, services or goods that they are looking for. A good taxonomy not only provides the basis for an effectively organised site, but it also increases the maintainability and adaptability of the site. Instead of a gradual build-up of haphazard links and degradation of the navigational structures, changes and improvements to the site can be done within a common, shared and managed information architecture.

Support for interoperability and integration

An increasingly important role for taxonomies is as a means to enable interoperability and integration at an organisational and an application level. Many governments have begun programmes to develop a common taxonomy and related standards for metadata as part of their modernisation programme. The same is true for organisations looking to link up their various applications, processes and knowledge bases in order to improve competitiveness and flexibility.

What is a corporate taxonomy?

A definition

A simple definition of a taxonomy is that it is a hierarchy of categories used to classify documents and other information. A corporate taxonomy is a way of representing the information available within an enterprise.

A classical taxonomy assumes that each element can only belong to one branch of the hierarchical tree. However, in a corporate environment such formal ordering is neither feasible nor desirable. For example, a document on a competitor's product may be of interest to different departments in the organisation for different reasons – forcing it into a single predefined category may be neater, but also reduces its usefulness. Corporate taxonomies need to be flexible and pragmatic as well as consistent.

An example

Figure 1 shows an example of a taxonomy. This is part of the UK government's standard category list for use on all UK public sector websites (for more information see www.govtalk.gov.uk). In fact, this is just an extract from a specification that runs to many pages.

Figure 1 **Extract from the UK's Government Category List (GCL)**

Crime, law, justice and rights

- Animal rights and welfare
  - Animal experimentation
  - Hunting
- Civil and human rights
- Crime
  - Antisocial behaviour and disorder
  - Arson
  - Business crime
  - Domestic violence
  - Drug-related crime
  - Murder
  - Offenders
  - Organised crime and terrorism
  - Racially-motivated crime
  - Sex offences
  - Smuggling
  - Theft and burglary
  - Vehicle crime
  - Victims of crime
  - Violence against the person
  - War crime
- Emergencies
  - Civil emergencies
  - Fire service
  - Police
- Ethical issues
- Extradition
- Firearms
- Justice system
- Law
- Security

Economics and finance

- Capital and financial markets
- Economic development
- Euro and EMU
- Investment
- Labour market
- Monopolies and mergers
- Nationalisation/privatisation
- Personal finance
- Public finance
- Tax
- UK economy

Government, politics and public administration

- Central government
- Civil service
- Constitution
- Devolved administrations
- Electoral system
- Honours system
- Local government
- Policy making
- Political parties
- Public administration
- Regional policy

As well as providing an example of a top-level taxonomy, the UK's GCL demonstrates some important points about taxonomies:

- this taxonomy does not stand alone – it is implemented in specific government sites, departments and services according to their needs, and is combined with other taxonomies relevant to those subject areas
- it does not dictate the technology to be used. Government departments and agencies are free to use a range of technologies to support the implementation of the taxonomy
- the taxonomy categorises the key concerns and issues that are the responsibility of, or involve, government agencies. It is not an organisational chart for the UK government – although it is inevitably influenced by the specific context of UK history and governance (some countries might not put hunting under 'criminal justice', a reflection of the highly contentious nature of this issue in the UK). Similarly, a corporate taxonomy should not simply be a map of current organisational structure or responsibilities, but should reflect the particular ethos and concerns of an organisation
- access and interoperability are key drivers in the creation of the taxonomy. It is part of a broader framework: the e-Government Metadata Standard (e-GMS). The e-GMS in turn is part of the overarching e-Government Interoperability Framework (eGIF). e-GIF lays out the UK government's policy and specifications for achieving interoperability and coherence across public sector information and communications technology. For any organisation, a taxonomy should sit within a broader strategy for information management with clear business objectives, and supported by associated initiatives on metadata standards, for example.

The role of technology

There is no intrinsic need for technology in the definition and management of a taxonomy. However, there is a growing role for tools that can assist or even eliminate some of the tasks associated with taxonomy design and management.

Tools and solutions are available that can assist with any stage of the process; from the simple editing and design of a taxonomy structure, to the automatic identification of categories and the assignment of content to the relevant classes.

Vendors disagree over the right techniques and approaches to the classification of information. The approach that suits any individual organisation will depend on a mixture of the business requirement, the type, volume and volatility of the information to be managed, the skills available in-house and the budget available.

The greater the volatility of the information and the categories to be used, and the less the in-house experience available, the more attractive an automated solution will be. For organisations with extensive experience of in-house taxonomy design, greater manual control may be preferred.

However, few cases will be simply 'either/or'. Even in an environment where information classification is largely automated someone still has responsibility for the

effectiveness and performance of the system (in terms of information access and not just technical reliability). For organisations with extensive knowledge of taxonomy design, increased automation can reduce maintenance costs, increase efficiency and extend the applicability of a taxonomy across an enterprise.

Building a corporate taxonomy

The preliminaries

Roles and responsibilities

Identifying who should 'own' the corporate taxonomy can be a difficult task for organisations that are new to such projects. In many large enterprises there will already be a corporate librarian, and their 'information science' experience will make them the natural candidate for this role.

Alternatively, an existing knowledge manager – or knowledge management team – will take on the responsibility of ownership.

If neither of these options already exists, the organisation will need to create a new role – and ensure that the appointee gets the training and, equally important, management support that they will need to be effective in what can be a politically-sensitive position.

In large organisations it will be sensible to split responsibility between different roles. For example, a knowledge manager may take on the task of liasing with users – understanding their requirements and also developing their commitment to the system – while an administrator takes responsibility for editing and managing the detailed elements of the taxonomy.

In most cases, a number of stakeholders (from the corporate library, the IT function, knowledge management team and business units) will need to be regularly involved in the ongoing management of the taxonomy and related standards.

Selecting a team

In order to create a corporate taxonomy, an organisation needs to define who uses what information. In order to do this it is vital that the project team comprises representatives from across the organisation – thus ensuring that the various uses for different types of information and documents are captured. If the process is undertaken using too small a group of people there is a danger that the taxonomy will be (unintentionally) biased. For example, the marketing group may not realise the value of a document to the product development team. In order to minimise such subjectivity it is vital to assemble a cross-organisational project team.

Due to the implications of such a large-scale project, it is also important that there is support from senior management. Implementing a taxonomy in an organisation will affect the way people work, and managers will need to be convinced that this is a positive move. Without top-level management agreement and championing of the project, it will be difficult to get support from all areas of the organisation.

The other group that must be closely involved is the users. Any testing or refining stages during the project need to include feedback from users. This will avoid the danger of users having a different perspective on what should belong in which category and therefore not being able to find information. A taxonomy that makes sense only to the project team and departmental managers, but is illogical in the eyes of the users, is useless.

The problem with creating a corporate taxonomy in-house, without support from external specialists, is that few people know how to do it, or have gone through the process before. However, there are usually people within the organisation who have a good understanding of the way in which their department or area would categorise information; these include website managers, content managers, librarians and database managers. Many organisations will be able to draw on a combination of these skills in order to establish an initial taxonomy.

Be realistic about the investment and resources required

It is important that organisations are realistic about the cost and investment needed to develop a corporate taxonomy. It can be a lengthy and expensive process, depending on the ambition of the organisation, and the required investment in time and money should not be underestimated.

Due to the nature of the taxonomy, and the need to satisfy every team or department in the organisation, it is not something that can simply be outsourced to a third party; it will require commitment and resources from all areas of the organisation.

In addition, once the taxonomy is in place there will still be costs to maintain it – ensuring that it continues to serve its purpose, and that it does not get out of date or fall into disuse. For this reason, the need for senior management support is vital.

One side effect of creating a corporate taxonomy is that it can prompt departments to rationalise and update their content, and improve their authoring processes. The obvious benefit of this is the superior quality of the resulting content. However, the downside is that this activity will hamper the progress of the project, as time is spent improving content rather than organising it. Any organisation looking to create a corporate taxonomy needs to be aware of this, and to take it into account when planning the project, to ensure that contingency for this is built into the process.

Seek help

Most organisations will benefit from outside expertise on the taxonomy project. The level of support required depends on the level of in-house skills available, but an external view on requirements and procedures can be invaluable.

It is also possible to kick-start the taxonomy itself by using standard taxonomy models and thesauri. Consultancies and software vendors are also able to offer template taxonomies to provide a good basis that can be tailored to specific needs.

Stay focused on the business need

The biggest risk to a corporate taxonomy project is that it becomes an end in itself and therefore detached from the original business case. We call this the 'enterprise data warehouse' syndrome following the tendency of organisations in the 1980s to embark on all-embracing data warehousing projects. Often, the effort to define the perfect warehouse model obscured the real business reasons for the project in the first place. Taxonomy projects that take this path will – like many data warehousing projects – be abandoned, unfinished or implemented late in an environment where the requirements have changed so drastically that the benefits are greatly reduced.

The taxonomy should be treated as a serious project worth an appropriate investment of time and resources, but it should not become an end in itself. It will never be perfect and it will never be completed – it is too closely tied to a dynamic and complex business environment for that to be possible. Keep the business goals in mind at all times, and aim to provide a taxonomy that can support those goals now, not in two or three years.

The process: creating a taxonomy and classifying resources

Carry out an information audit

Before the project team can start identifying categories by which to classify the organisation's information, it will need to decide on and locate the resources to be included.

The starting point may be commonly accessible applications on the corporate network. There will also be relevant local systems that contain information that will not be visible at a corporate level. Access issues surrounding these systems will need to be clarified prior to the project.

It is also important to get a sense of the likely volatility of the taxonomy during the planning and design phases – this will help determine the future governance model and support costs.

Once the organisation has a good overview of the information that needs to be classified, whether through the use of analysis tools or through combining the knowledge of the various people in the project team, it will need to establish the structure of the taxonomy.

Using an existing taxonomy or other categorisation models

Any existing taxonomy structure is an obvious starting point. Even partial or incomplete taxonomies can help the project team to understand current usage and to determine what metadata is already available to aid classification.

However, organisations should be very cautious when using existing departmental or file structure taxonomies as a starting point. These tend to give a tunnel view of the actual information due to the purpose of the particular department that has previously used them. Also, because they are constantly growing and rarely rationalised, they frequently give an out-of-date or duplicated view of corporate information.

You should also be wary of using an organisational structure as the basis for a taxonomy – consciously or unconsciously. While organisational structure is an important aspect of information use (and therefore of taxonomy design), the goal should be to think outside such political boundaries.

Also, look at the viability of using third-party taxonomies. As well as standard taxonomies developed for areas such as medicine, defence and government, consultancies and software houses can provide base taxonomies relevant to specific industries (at a cost).

Using categorisation technology to define a first-cut taxonomy

It is possible to define and build a taxonomy manually – and there are various techniques that can be used, such as card files and structured workshops, to achieve the necessary input. For many organisations, making use of software tools is an effective means of reducing development time and supplementing a lack of in-house skills. A range of tools is available for taxonomy design, document tagging and taxonomy management, but the most important tools are those for categorisation.

Most categorisation tools require some initial structure and content, which they analyse in order to build a model for classification. The initial ‘training set’ may be a fully defined existing taxonomy or a set of directories providing some basic organisation of material. The model can then be tweaked by manually changing weightings and assignments or by extending or changing the training set.

Some tools can produce an initial taxonomic order from an undifferentiated set of information. However, in this case you will still want to closely review the proposed model to make sure it fits your corporate view of the world.

Fully-automated categorisation is very useful for organisations with no experience in building or designing a taxonomy (but cannot be used as a way of avoiding ownership of information needs). It is also useful where the information is highly dynamic and a quick but non-rigorous sorting of information is all that is required.

Refining the taxonomy

When creating the hierarchy the organisation should ensure that, while it is important to represent the knowledge base, the purpose of creating the taxonomy is not forgotten. It must be logical and navigable, and it should always be clear to users, as well as to the project team, why a certain document or piece of information falls into a particular category (or group of categories).

A significant challenge when defining a taxonomy is to create a balance between the breadth and the depth of categories. If categories are too specific in their description and classification there is a danger that they will be too transient, and need changing as their contents change. On the other hand, if a category is too broad it will hinder navigation as the precise nature of its contents will be difficult for the user to determine. A taxonomy can theoretically have an unlimited number of levels; however, too many will mean that the user becomes lost navigating down to the bottom level. A narrow taxonomy also forces uses to make a navigation decision with too little information and to have to work through too many layers before knowing if they are following the right path. The general consensus is that breadth is better than depth as far as usability is concerned.

The project team should establish a set of guidelines to determine, for example, the maximum number of layers in the hierarchy structure, the maximum number of documents per category and the maximum number of categories per document. These numbers need not imply implacable limits, but such guidelines will help the project team, and the person who is responsible for maintaining the taxonomy, to determine whether or when a single category needs to be divided into further subcategories.

Testing

Once the project team has created a taxonomy structure that represents the corporate knowledge it is important to test it. This means that it is necessary to classify a sample selection of documents from the knowledge base to determine whether they fit well into the structure, and whether all of their possible purposes are covered by the taxonomy. It is also necessary for users to test the structure with those documents in place, to see whether the user's expectations of what should fall into that category match the taxonomy, and to see how easily the user can navigate through the layers of the hierarchy.

Once this testing has been done the project team can refine the structure; renaming any categories to ensure they are meaningful and consistent. Testing and refinement is an iterative process.

Applying the classification model

The project team needs to categorise the remainder of the organisation's documents once the taxonomy has been tested and finalised.

Automatic document categorisation tools need to be trained against a set of sample documents for each branch in the taxonomy hierarchy.

The training process needs to be monitored, and training sets will often need to be tweaked or extended to get the best results. Many products provide feedback on why a document is assigned to a particular category (for example, the keywords that have influenced the decision), or they provide a confidence factor that indicates the level of certainty involved – this information can help the administrator to refine the classification process.

Monitoring

Classification of documents and information is necessarily an ongoing operation: documents will need to be categorised as they are created. An automatic classification tool will be able to categorise documents as they enter the knowledge base, thus reducing the responsibility that is placed upon the document author.

Corporate taxonomies are not static in their nature – even the best taxonomy will need updating regularly in order to ensure that documents are still being categorised in a logical and comprehensible manner, and that the categories are still useful and current. There is therefore a requirement for an ongoing budget and resources in order to maintain the taxonomy.

The evolving role of software: beyond categorisation

Software support for the corporate taxonomy is not limited to the provision of automatic classification tools. More effort is now being invested in software that can increase the usability of taxonomies for both corporate users and consumers and in tools to support taxonomy design and management.

Multi-faceted taxonomies

The limitations of a purely hierarchical taxonomy model have been recognised for many years in academic and information science circles. Consequently, there has been considerable interest in taxonomy structures that offer a more flexible view of how information can be categorised for general use: these alternative approaches are often referred to as faceted, multidimensional or relational taxonomies. These concepts are now making their way into commercial products aimed at supporting taxonomies in an e-commerce or corporate environment.

Multi-faceted taxonomies enable the user to navigate through a number of facets of a taxonomy (for example, by artist, genre, instrument or composer in a music library). They also allow the different facets to be cross-referenced to narrow or widen a search as the user browses the categories (for example, you can browse and select recipes by a combination of ingredients, cuisine and occasion at www.epicurious.com).

Developments in multi-faceted taxonomies are also closely linked to new analytical and visualisation capabilities that offer to transform our experience of search and navigation through large volumes of information.

(For an academic view on the advantages of multi-faceted approaches to taxonomies and metadata see the Berkeley University Flamingo project at http://bailando.sims.berkeley.edu.)

Workflow and collaboration

Developing and managing a taxonomy is a collaborative project involving multiple stakeholders. It also needs clear procedures for change management. Integrated workflow tools and collaborative editing tools make it easier to manage taxonomies in large organisations and places where taxonomies have to monitored and adapted on a regular basis, such as shopping sites.

Search analytics and taxonomy management

Search analytics refers to the collection, analysis and exploitation of information about the way search technologies are used. The initial driver for this development came from the need for e-commerce sites to know how users are searching their site. The next step is for these techniques to be used within the enterprise. Better information

on what users are searching for, and the ability to tailor results and navigation paths, offers a relatively easy way to improve information retrieval within an organisation. There is a great opportunity for using search analytics in the design and maintenance of better taxonomy structures.

Visualisation tools

Improved visualisation capabilities can enhance the value of taxonomies at two levels:

- usability – providing visualisation capabilities to the user enhances their ability to take advantage of the investment in an underlying taxonomy. Taxonomies provide a basis for implementing existing visualisation tools in a useful way and open the way for new tools that can help users visualise the multidimensional space in which they are searching
- design and management – improved means of visualising a taxonomy structure make it easier to ensure an efficient balance amongst categories and better fit with user expectations. Such developments are closely linked to improved support for the rapid design, test and refinement of taxonomies.