Challenges, successes and failures in automating taxonomy management

online information 2003

Daniel Brown

Deputy Chairman, APR Smartlogik, Cambridge, UK

Introduction

Unstructured data is difficult to manage, especially when it is scattered among different formats from emails to Powerpoints to Word documents. Navigation is difficult, often impossible. All too frequently in public and commercial organisations information is totally inaccessible both for employees and for customers/citizens who need to use it.

If the user has to rely on searching alone, then all the effort in locating information is down to them, and they do not know the contents of the repository. Structuring the information in advance assists the user before they even begin their search.

What a taxonomy is

A taxonomy is a collection of topics or concepts organised to show how they interrelate. Hierarchical relationships (e.g. 'parent-child') and associative relationships (e.g. 'see also') capture the essence of a subject area by being described by the taxonomy. A taxonomy enables documents (in the widest sense) to be classified and organised according to a structure. It therefore enables users to access documents more easily and also guides them in exploring a document collection by laying out the conceptual structure of the subject area. Thus a taxonomy makes it possible to describe a plant so that in addition to what it looks like and where it grows we can see what is most closely related to and how it relates to other plants. There are different types of taxonomy. In the Linnaean system for naming animal species, for example, an item has only one parent in the hierarchy. Another example is SNOMED which makes a lot associative links to aid medical diagnosis (this problem is related to this drug which is in this category of drugs which has this side effect...). Other taxonomies do not have this singularity of parenthood – in a job description taxonomy a person might be both a novelist and a footballer.

A thesaurus is the logical compliment to a taxonomy. Where the latter contains only the name of each topic (called a Preferred Term) a thesaurus describes other names for the same topic (called Non-Preferred Terms). So, for example, a taxonomy might contain the topic 'rivers' and a thesaurus would tell us that 'streams', 'brooks' and 'canals' were other names for that topic. Generally we use the names taxonomy and thesaurus interchangeably to refer to the combination of the two.

Creating a taxonomy

Creating the taxonomy structure is hard work but a well designed one will pay a big dividend for the end user. Creating taxonomies is a mix of science and art and can involve either humans or technology or both. We must ensure the following:

1. The taxonomy should be simple and flat so that it can be navigated easily. If it becomes too deep and complex then users will end up lost.
2. There is latitude for the taxonomy to grow, adapt and change over the lifetime of its use.
3. The audience of all users (both internal and external) and what they want from the system is understood. In particular, the taxonomy should cover the topics of interest to its audience.
4. We understand that the dynamics of the taxonomy and the information it describes are different. The organisation's information content may change rapidly, but the knowledge structure should be sophisticated enough to contain these alterations without much change itself.
5. The taxononomy should (a) conform to ISO or ANSI standards; (b) contain scope and historical notes or other relevant metadata for each Preferred Term to explain its usage and how it came to be in the taxonomy; (c) have restrictions over users adding or removing terms

It is usual in presentations to say that the key decision is to whether to use an automated taxonomy generation tool or embark on a laborious manual process of devising the taxonomy yourself. Unfortunately, this dilemma

usually derives from over-eager sales people in software companies. Whilst we have reached the stage where automated tools can assist in the construction of a taxonomy, I do not believe that it is profitable to rely solely on such tools for manufacturing a taxonomy structure from an existing set of documents.

The reasons for this are varied and subtle enough to intrinsically illustrate the challenges quite well. To create a taxonomy we need to understand a complex hierarchical network of relationships between concepts. Computers can struggle where humans have a natural aptitude. Thus a referral word (e.g. 'wildlife') may not itself even be mentioned in a single text which it characterises. Furthermore, there is a requirement that such a characteristic word is *understood by users* as referring to a domain of texts. Such understanding requires a set of capabilities including contemporisation (i.e. certain terms are more in vogue at certain times than others) and localisation (idiomatic use by region). These capabilities are currently beyond the scope of most technologies and we therefore require human intervention. This implies a minimum design of knowledge architecture using an automated thesaurus builder in conjunction with human consideration.

On the positive side, we usually find that for a business or organisation rich in information, this stage of taxonomy creation will already have been partly completed anyway. Some form of structure for organising information is essential simply to manage the data. So the human work requirement is not usually perceived as too much of a burden. Furthermore, there are benefits in working with a structure that people already know. Consider a taxonomy design for the field of science. From our life experience alone it will look something like:

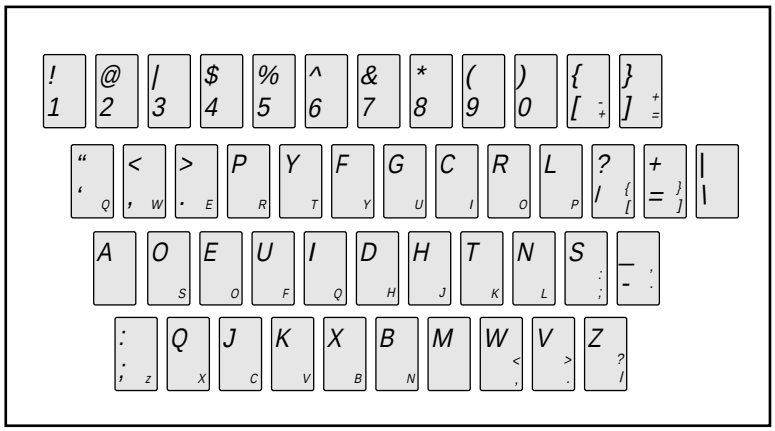
*science

* chemistry * physics

* node * node * node *node *node *node

It makes far more sense to utilise the taxonomy structure that we are familiar with because it is just this familiarity that provides us with the ability to use it efficiently. While a QWERTY keyboard may (deliberately) not provide the most efficient design for finger manipulation, it may nevertheless be speedier for the majority of users than an inherently superior DVORAK keyboard design. Equally, a poorly designed but familiar taxonomy structure is likely to be more efficient than its theoretical perfected counterpart because of familiarity.

DVORAK keyboard



Computers cannot account for this, because they are not equipped to know what is familiar to most people. Consequently, the best way to proceed in taxonomy design is to manually build an overarching structure using preferred terms that are familiar to users. To some extent this process is independent of the document corpus. We can then *augment* this using computer software to scan the existing corpus of documents and determine some of the lower more specialist child terms that should be used (e.g. in the example above we manually create the science, physics/chemistry structure and the software automatically helps determine what terms to use in characterising physics-related documents).

Making the link

Once we have the taxonomy, we can begin labelling documents either manually or automatically. We should bear in mind that humans are slow, inconsistent but ingenious. They can also vary from specialist information scientists to bored underpaid school-leavers. On the other hand software tools are fast, consistent and meticulous. In order to assess the capabilities of automatic classifiers we need to evaluate:

1. Precision and recall.

2. Threshold where for automatic classifiers borderline documents are regarded as uncertain.
3. Building indexing knowledge for parent and child terms.
4. How to include subtle human expertise within the system.
5. How fast and efficient the system is as a complete process.
6. Corpus quality – source information for taxonomy.
7. Arrangement of taxonomy classes – there should not be more than five levels and the following qualities should be noted: (a) depth: the average depth of a sub-tree from a given node; (b) width: the number of children from each node.
8. Terminology – terms used to name classes of taxonomy.

The requirement is then to label documents with the appropriate metadata. To do this, a range of methods exist including:

1. Bayesian classification.
2. Automatically generated rules (with training set).
3. Automatically generated rules (from terms in taxonomy).
4. Manually tweaked automatically generated rules.
5. Manually generated rules bases.

The decision for which methods to use can be made using a set of rules (heuristic). A reasonably effective heuristic is:

1. Assess how many documents there are in the categorisation concept and set a minimum of n (typically n is at least 100). If there are at least 100 documents then we can use a Bayesian classifier.
2. If the number of documents (ND)< (n/2) then we use manual classification.
3. If 100>ND<(n/2) then we can use automated rules base generation.
4. We can later examine the rules base and decide which ones we might want to tweak – typically this will be done according to which node of the taxonomy is most important. For example in a system which automatically scans news in order to refer stories to the PR department, we might determine that the company itself is so important that the rules base should be manually augmented.

Table 1. Classification Technology Chart

	Requires training set	Precision	Recall	Best suited for
Bayesian	Yes	Poor	Strong	Fuzzy concepts that are hard to define and already exist
Automated rulesbase (from training set)	Yes	Fair	Poor	Rapid accurate sorting of clearly defined concepts
Automated rulesbase (from taxonomy terms)	No	Good	Fair	Well developed taxonomies with many synonyms
Manually tweaked rulesbase	Yes	Good	Good	Accurate sorting of well defined and unusual concepts
Manual rulesbase	No	Outstanding	Good	Exceptional accuracy and idiosyncratic concepts

Maintaining the taxonomy

This process is often underrated and not accounted for in budgets. Whilst a taxonomy should not require radical change it will need guidance and planning in order to ensure continued relevance (potentially for new groups of users). Maintenance issues can include:

1. One taxonomy term is overloaded and should be cut into ‘siblings’ or a parent and children.
2. A term has become outdated (but the position in the taxonomy remains the same).
3. A new concept emerges and should be inserted into the structure with a meaningful name.
4. The demotic language describing a term (its synonyms) changes over time. For example, during the US election of 2000 the term ‘Chad’ became synonymous with the US electoral process, whereas previously it had described the African country.

Case Study: UK Government's online strategy

The UK Government's online strategy aims to create compliance within an e-government interoperability framework (e-GIF) directed to the public sector. The e-GIF sets a series of standards and policies and aims to make it easier to manage data which in turn should make it easier to find information and services. The main objectives are:

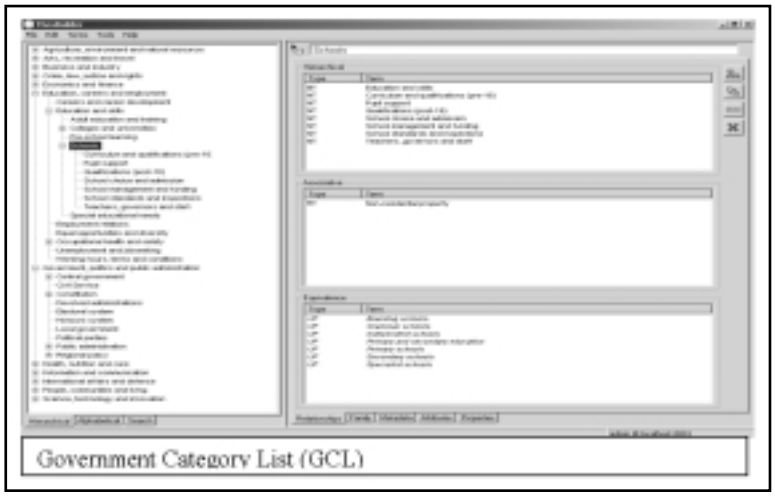
1. Improving management of official information that will make government more efficient and ensure it meets obligations for handling public records.
2. Transparent access to information without having to know which government organisation is responsible.
3. Providing access to multiple information sources from a single web portal.

A proprietary 'Semaphore' software package (developed by Smartlogik) enables users to build and refine taxonomies with logic conforming to the ISO guidelines on taxonomy construction ensuring that the taxonomy's integrity is maintained (e.g. duplicate preferred terms and excluded child terms are automatically updated on the deletion of a parent etc.). The system can handle complex, multi-hierarchical taxonomies of tens of thousands of terms, a task many people struggle to accomplish using spreadsheets.

The scope of the e-GIF encompasses any citizen-facing interface, any system where data is sent to another public sector organisation and any intra-government information services. To achieve these objectives, compliance with e-government standard (e-GMS) is required. This ensures that, amongst others, a common Date, Title, Creator and Subject tags are attached to every document. By applying tags, documents can be collated in a consistent manner.

However, creating a common placeholder in a document is of little use if everyone adds their own content. This is especially true of the Subject metadata. To ensure that a common and consistent language is used in the document metadata, the Office of the e-envoy has created the Government Category List (GCL). This is a taxonomy of unique terms to be used with the Subject Category element of the e-GMS. For example, a document about freedom of speech and censorship should be tagged with the GCL category 'Civil and human rights'. To conform to the e-GMS tagging, information with at least one valid heading from the GCL is mandatory.

The e-GMS recognises that the GCL is limited to high-level headings and benefits from the addition of more specific terms from a vocabulary designed for a sector. In this case, other vocabularies more relevant to the subject area may be used in addition to the GCL.

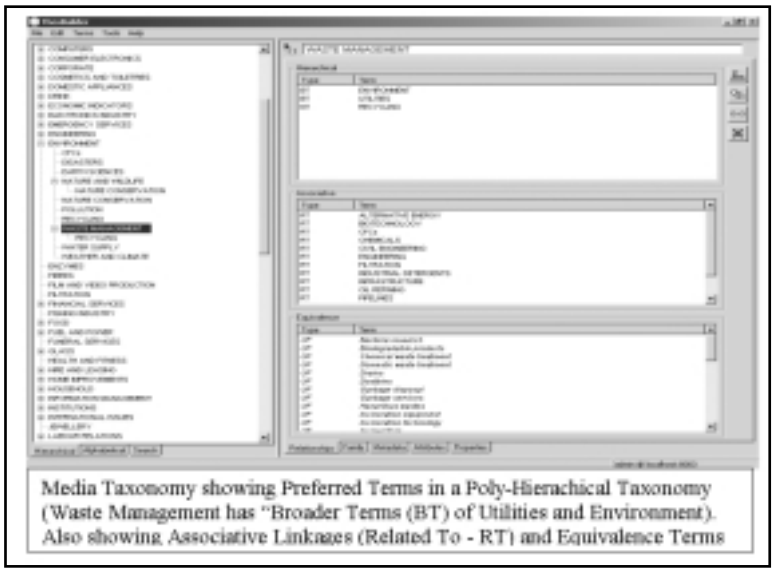


One such vocabulary, known as the APLAWS Category List (ACL), has been created by the Local Authority Websites National Project (LAWS). Coordinated by West Sussex County council in the UK, with other partners including Camden, Sutherland, Tameside and five 'Welland' partnership councils, this project has a number of workstreams including transaction software for managing contact with citizens, secure messaging and templates for publishing and updating content.

Its information architecture stream has held a series of workshops with the aim of providing a deeper set of subject terms relating to council services, such as those often seen on a council's A-Z of Services list ('abandoned cars', 'recycling', 'street lighting', etc.). Councils adopting this common category list will again find it easier to share information internally within e-partnership initiatives and in providing citizens with access to their services.

Most local authorities are endeavouring to become e-GMS compliant implementing content management systems (CMS) to control content authoring and enforce metadata attachment. However this neither solves the challenges of tagging historic data nor facilitates citizen access to the newly created information. The Smartlogik Semaphore system enhances existing Content and Record management initiatives and websites by managing, extending and exploiting the subject metadata. Based on existing Category lists, the system provides the capabilities to refine taxonomies to local requirements and employ them for automatic document classification and intelligent document search.

The system is important to end users because of the necessity to maintain a consistent set of categories (or Preferred Terms) applied to the subject metadata. However, the language used to describe these unique categories (i.e. Non-preferred Terms) can vary widely across the country. For example the preferred term 'Comprehensive Schools' might best be described on a local website by naming the actual schools of that type in the district. Moreover in many areas of the UK local idioms are how citizens will refer to services and these local phrases must be added to enrich the taxonomy. A taxonomy management module maintains and extends existing taxonomies. The evidence to extend the thesauri may be found in the search logs from a public website (listing the actual search terms input by site visitors), or a collection of meeting minutes already classified to a particular topic.



Media Taxonomy showing Preferred Terms in a Poly-Hierarchical Taxonomy (Waste Management has "Broader Terms (BT) of Utilities and Environment). Also showing Associative Linkages (Related To - RT) and Equivalence Terms

In order to achieve the high precision and recall indicated in the table, it was clear that we should use the automated rulesbase from taxonomy terms. The Structure Classification component uses the language held within the taxonomy to determine if a document is associated with a Preferred Term. It uses a complex set of rules to determining the importance of words and phrases occurring in a document and if enough of the taxonomy language is found, the relevant category tags are assigned. This approach ensures that:

1 Metadata is applied consistently and accurately.

2. An objective view is applied that correctly applies multiple category tags.

3. Large historic archives of information can be automatically and rapidly tagged without the substantial cost overhead of employing people to read and manually classify them.

4. The tagging can be built into the CMS workflow process to suggest appropriate tags for inclusion.

Metadata from multiple taxonomies can be automatically included in the document, e.g. GCL category 'Sports and Recreation'. Local Council specific categories include 'Sobell leisure centre', 'Learn to swim scheme'.

If no obvious category is found by Structure Classification module, a definite 'not tagged' category can be returned (if for example the file is written in a foreign language or a binary format). Identifying these exceptions can allow workflow to be developed that flags documents to the users for manual classification.

In summary, the system enables an expandable knowledge asset based on a taxonomy (such as GCL and NCL) and a method for automatically tagging documents with the subject metadata. The technology adds value to existing investment in content or document management systems. The taxonomy provides an extremely valuable knowledge asset for an organisation that is useful in its own right but has additional power when applied to classification and search applications. The system has successfully improved citizen access to information and increased online 'self help' whilst reducing expensive load on call centres and it has removed the requirement for teams of people to tag documents.

It is important to note that this system does not directly make search decisions for an individual; rather, it provides a decision subset from which a human can make the final decision. In this, as in many other examples, we are still a very long way away from the technology actually making the final decision. Here, as elsewhere, the technology provides imperfect recommendations which can quickly be dismissed by a human, along with other recommendations which provide significant value, and the balance of precision and recall should be such that the precision is not too low for the recommendations to be worthless and the recall is equally not too low for many items to be left out.

However, I believe that even these imperfect recommendations are enough to convince users that the search engine is 'intelligent'. Once users accept this, they will be more willing to engage with the search to navigate to the information they want, as opposed to inserting single search terms and giving up when the required information is not in the first three results.

Conclusions

1. Taxonomies are providing us with opportunities for wider, more productive access to information.
2. There is a significant need for assistance in automated taxonomy management, but systems cannot yet automatically generate the taxonomy alone.
3. Automated assistance in taxonomy generation is already effective and methods are improving.
4. Requirement is to create a structure through general terms descriptive of a range of themes from a range of diverse documents that may not be structured in the first instance.
5. Practical benefits are already being achieved.

Contact

Daniel Brown
APR Smartlogik
160 Euston Road
London, NW1 2LZ
UK

daniel.brown@aprsmartlogik.com
www.aprsmartlogik.com