Planning Controlled Vocabularies for the UK Public Sector

Abstract: In the UK, the aim to make public sector information much more available to the citizen has led to establishment of an “e-Government Interoperability Framework” based on a set of core standards. Among the standards is a controlled vocabulary, known as the Government Category List (GCL), used to select keywords for the metadata of all electronic resources originating from central or local government. The GCL is a small and simple taxonomy, designed to facilitate high-level browsing rather than deep searching. Specialized thesauri for particular subject areas may optionally complement the GCL. To ease the indexing burden, GCL terms will often be selected by direct mapping from the specialized vocabularies.

1. Introduction

Today almost all governments are feeling both the pressure and the opportunity to make their own information freely available to the public. The opportunity extends beyond provision of information to enabling electronic delivery of services, such as the filing of tax returns or applications for passports. Furthermore, there is widespread recognition that it is not enough just to release large swathes of documentation on yet another website. Information should be organized in such a way that it will be easy for people to find.

But how to achieve this? In the UK, one thrust is to set standards for “interoperability”, basing them on existing ones that are non-proprietary and already widely used, such as HTTP, XML and PDF. Additionally a new e-Government Metadata Standard has been prepared, based on Dublin Core and developed in partnership with the Dublin Core Metadata Initiative (http://purl.oclc.org(metadata/dublin_core/) and the governments of other countries. The Office of the e-Envoy (www.e-envoy.gov.uk) is leading UK developments.

The dream is that one day information and resources will flow seamlessly within and between all public sector bodies. The citizen at just one port of call will be able to find what he wants, from anywhere within officialdom. And one of the ways he or she ought to be able to find things is by subject. The present paper is an account of the debate and decisions on what to do about access by subject. In particular, given a “Subject” element in the new metadata standard, how should it be filled? At an early stage it was felt a thesaurus should be used for selection of subject metadata, but the question then arose, “What sort of thesaurus?”

2. Thesaurus options

A huge confusion surrounds the word “thesaurus”, which some people use to mean a controlled vocabulary conforming to ISO 2788 (International Organization for Standardization, 1986), and others imagine as a sort of electronic Roget model, automatically swapping synonyms behind the scenes. For this project, six different options were considered, as well as sundry variations on them:
Models A and B above were both seen as controlled vocabularies, in that each would comprise a finite set of terms to be used for indexing as well as retrieval. The main difference between them is in size and complexity. Model A would comprise 7000 or more terms, to cover the full span of public sector interests, with inter-relationships conforming to ISO 2788. Model B would be much smaller, with perhaps 400 preferred terms and plenty of non-preferred lead-in entries. It would have limited functionality, designed for high-level browsing and navigation rather than precision searching. Its relative simplicity would make it easier for indexers to use.

Model C is an acknowledgement that many government departments and agencies already have, or are developing, their own controlled vocabularies designed for their own sectors and contexts. A thesaurus confined to one subject area should give much better results than one aspiring to cover the public sector universe. So maybe the development of specialized thesauri should be encouraged, and central government should build a “meta-thesaurus” integrating all the vocabularies into one reference tool that shows linkages between related concepts in the different vocabularies.

Model D avoids the expense of indexing or meta-tagging at source. A search thesaurus is used only at the point of search, and the whole hassle of indexing disappears. Although there is no standard for search thesauri, several examples do exist, such as the Contemporary Thesaurus of Social Science Terms and Synonyms (Knapp, 1993), a prototype proposed for a pharmaceutical firm (Lykke Nielsen, 1998) or the synonym sets built into some search engines. The function of the synonym sets is to expand a user’s query by adding synonyms, usually strung together with OR logic. (Dextre Clarke, 2000). It is possible also to adapt an ISO-2788-style thesaurus to this type of use, exploiting the equivalence relationships for search expansion.

Model E abdicates responsibility for a pan-government vocabulary, but encourages information professionals in diverse agencies to share ideas on building and implementing specialized thesauri for their own sectors. This is certainly the cheapest option when seen from the input viewpoint, but in a cross-sectoral context does little to mitigate the end-user’s expenditure of time on fruitless searching.

Finally, Model F is a more positive move towards relying on technology rather than human indexing. A taxonomy or classification scheme would be built and would be presented to the users of portals. But automatic categorization would remove the need for human indexing. Human intervention might still be needed to define and tune the categorization rules and algorithms, but this would be very much less expensive than requiring an indexer to process every document at source.
3. Workshop discussion

If any type of controlled vocabulary is to be used for assigning metadata when documents are first input, it is essential to have the cooperation of those on whom the task will fall. There is a big contrast between coordinating document input procedures in one organization, and procuring the compliance of all the bodies in the public sector! The Office of the e-Envoy therefore held a workshop of key stakeholders, with 28 government departments and agencies represented. The above models were discussed, and the implications for implementing them.

The first thing noted about the options was that they are not mutually exclusive. It would be possible, perhaps even recommendable, to opt for more than one. Some of the models could be adopted unilaterally, either sooner or later. But with Models A and B it would be important to make an early decision because they need concerted action from all parties.

Most participants in the Workshop had an information management background and a conviction that Model A was the one most capable of yielding high-quality search results. However, they were also aware that consistent indexing requires careful quality management, and is best handled by a small team of trained staff. The reality in offices today is that web "publishing" is often carried out by the authors of the original documents, or by a webmaster who is hopelessly under-resourced for the meticulous task of adding metadata. Rarely have they been trained in indexing (often called "meta-tagging" in this context) and even more rarely in how to use a thesaurus. If a significant proportion of all the resources to be searched have inconsistent subject metatags, randomly applied or omitted, then searches will yield poor quality results and Model A falls into disrepute. When the need for training searchers was added to this scenario, Model A looked even less practical.

Model B was thought more feasible. If the controlled vocabulary is presented as an expandable hierarchy with a small number of top terms, and if just 2 or 3 mouse clicks are enough to find and select the most appropriate heading, then reasonable quality can be achieved for both indexing and retrieval. Such a simple hierarchy will not of itself enable precision searching, but is good for high-level browsing.

Model C, the meta-thesaurus, is much more ambitious in terms of the building and maintenance effort. A survey of the closest central government departments revealed that 23 specialized vocabularies were already in existence, without counting all the others in local authorities and executive agencies. Specialized software would have to be commissioned to handle such a large and complex volume of data, and a team of experts would be kept busy for years.

The search thesaurus, Model D, requires less investment and could be a promising approach. However, an immediate decision was judged unnecessary because no provision needs to be made for it in the metadata standard, and no human indexing is required. The option is still open for the future.

Model E was popular with all participants. Thesaurus building is for specialists, who have much to gain from comparing notes with others in the same business. In some sectors several organizations have overlapping interests and could well share the same specialist vocabulary. The discussion forum may be physical or electronic or probably both.

Model F is much disparaged in some quarters, but there are circumstances where human indexing is simply not available and automatic categorization is the only hope. This occurs, for example, with legacy databases or when external
resources gleaned from the World Wide Web are to be integrated with internal collections. Categorization technology is improving all the time. However, the structure, format and syntax of taxonomies driving automatic categorization tend to be specific to particular proprietary software suites. The implementers of particular portals may appropriately handle development, unilaterally without need of a government standard. So the Workshop left this, too, as an open option.

4. GCL description

The clear choice to emerge from the Workshop was Model B, the high level taxonomy. It has been named the GCL (Government Category List) and now forms part of the mandatory metadata standard. Government departments, executive agencies, local authorities and other public bodies are now developing the most cost-effective way of adding metadata to all their electronic resources.

Mandatory at the input level does not mean mandatory for display purposes. Portals implementing the taxonomy can choose between many approaches, of which one of may be seen on the Quick Find page of UK Online (http://www.ukonline.gov.uk/). Other websites will no doubt customize the display for their own target audiences.

As recommended by the Workshop, the GCL aims to be small and simple enough so that a reasonable level of consistency can be expected from the hundreds of people who will be obliged to do meta-tagging. At the time of writing it has around 360 preferred terms and over 1000 lead-in terms. The preferred terms are arranged in just 12 hierarchies, so that people browsing from the top down can easily select the most appropriate heading. Figure 1 shows the top level headings, and Figure 2 shows the full extent of the hierarchy below one of the top headings. For each of these headings the full record gives any scope note, lead-in terms and cross-references. We call the GCL a taxonomy because it is a hybrid between a thesaurus and a classification scheme, not conforming to the strict conventions required for either of these, but aiming to be user-friendly in an electronic environment.
An important aim for the GCL is to serve the citizen. The choice of terminology and structure is guided neither by academic "correctness" nor by the priorities of government employees, but by where ordinary people might expect to find things. The average citizen cannot be expected to know which government department handles which transactions, and so the GCL hierarchies studiously refuse to reflect the way UK government is currently structured.

Of course it is a huge challenge to find a structure to suit the way every ordinary citizen thinks. We all know that different people think differently. Even the same person views things from different perspectives on different occasions. Therefore the GCL is polyhierarchical, allowing more than one hierarchical access route to the same subject. And copious lead-in entries and cross-references are there to support people who just cannot find what they want by browsing.

While the taxonomy does not conform to ISO2788 guidelines for term relationships, in practical terms it is maintained using a standard thesaurus management package. It is available for download from the GovTalk website (http://www.govtalk.gov.uk/) in RTF or PDF format or as a zipped set of HTML files. The HTML version will run on almost any PC and offers a very convenient way of navigating the stand-alone taxonomy.

Suggestions for more terms, synonyms and relationships are still coming in. Realistically, they will never stop coming in. Following release of Version 1.0 in January 2002, updates are expected in May and again in September 2002. Consultation procedures are in place to help us steer clear of pitfalls along the way. The GCL will have to evolve to meet changing needs and expectations.
5. Distributed model

Since the taxonomy enables browsing rather than searching, it needs to be complemented by other tools and facilities. As already noted, many government departments and agencies already possess their own controlled vocabularies designed for their own sectors and contexts. For portals and websites in a particular sector, these specialized vocabularies may provide search tools well matched to the needs. But a big challenge confronts any agency using its own vocabulary as well as the GCL.

At the time of input, hard-pressed staff are unlikely to want to the job of indexing once with their specialized vocabulary and then again with the GCL. Therefore it is important for each public body to develop automatic mapping techniques. Typically, a mapping table is prepared showing, for each specialized term, the closest corresponding GCL heading. Then when that term is selected as a metatag, the GCL heading is automatically added too. Because the GCL is limited to high-level headings, the mapping is almost always from the particular to the general (a much more feasible proposition than the reverse process!).

Given the coexistence of specialized vocabularies and the GCL, it is interesting to compare the chosen approach with the meta-thesaurus model. To achieve the latter, it would be necessary to import all the thesauri, that is to say, all their terms and relationships, into one compilation. Considerable complexity arises when one term has different meanings in different vocabularies, or different relationships, or is non-preferred instead of preferred. With the current approach, all those differences can be tolerated, because they are not managed centrally. The sum of all the mapping tables could be likened to a distributed meta-thesaurus. But because the vocabularies and tables are managed separately, the complex
management issues never have to be addressed. The parallel is not exact, because the mapping tables are currently used only at the time of input, not at the time of searching. But in time, ingenuity could well find a way of exploiting assemblies of mapping tables as a search aid too.

6. Conclusion

We still have a long way to go before those seamless flows of information are achieved. But increasing proportions of the population are now sitting in front of a screen wired to the Internet, expecting to be able to find what they need. The GCL is just one small piece in a jigsaw that needs to be assembled for them. The implementation of standards will be vital in fitting the pieces together. Some pieces are already in place, giving service in particular areas. Others will be added progressively, and the picture that takes shape on the face of the jigsaw will evolve to meet changing needs and exploit new technology.

Two or three years may be needed before enough organizations have applied the GCL to enough resources so that we can evaluate the effects. At that stage it will be useful to revisit the debate on “what sort of thesaurus”, and see what more of the open options should be implemented.

References