Vocabulary Management at the Department for Education

Helen Challinor

helen.challinor@education.gov.uk

Helen Challinor is the Departmental Taxonomist and Information Standards Librarian at the Department for Education. She has 25 years' experience of working in government libraries in a range of roles from enquiry services to thesaurus management.

Summary

This case study considers the use made of controlled vocabularies at the Department for Education. It outlines the principles of vocabulary management, before explaining the uses made of controlled vocabularies within the department. It includes explanations about why decisions were taken, how users were engaged and a forward look.

Introduction

A controlled vocabulary, in its broadest sense, means different things to different people. It could be anywhere on a spectrum from a flat (non-hierarchical), alphabetical list, to an ontology with complex semantic relationships and all points in between.

Here the focus will be on the role played by a master controlled vocabulary file, from which we derive a hierarchical subject taxonomy and many flat pick lists. Together these drive a number of IT systems within the Department for Education (DfE). By using a single controlled master, we are able to provide a consistent approach to managing our information in these systems.

Controlled vocabularies are governed by standards and rules, but there are decisions to be made as projects evolve. These decisions have long lasting implications. How and why we made them is just as important as determining the subject content of the taxonomy.

User engagement is another important area in the development of any subject taxonomy. Every member of staff has to interact with our terms to a greater or lesser extent. This means that it is essential to have clear vocabulary structures in place and good support available, should it be needed.

Principles of vocabulary management

The master vocabulary file has been developed using an industry standard thesaurus management tool called MultiTes. Our instance of MultiTes holds a subject thesaurus and a series of authority files. A coding scheme has been devised using the categories functionality, which allows us to separate and extract the different types of vocabulary that we have designed.

The principles of thesaurus construction have developed over a number of years and have been codified and standardised in ISO 25964. The discipline of following these rules is helpful both to the manager developing the taxonomy, and in explaining it to users of the systems.

Ad hoc vocabularies created on the back of the proverbial napkin rarely provide a consistent framework. However, this is traditionally the approach adopted at the last minute to populate newly developed IT systems. Most developers have no idea that there is a science underpinning the management of subject vocabularies, and why would they? Neither do end users, and why would they? Devising innovative ways of explaining the nuances of vocabulary construction to developers and end users has been challenging, enjoyable and a significant deliverable from the project.

So, what is a subject thesaurus?

There are five elements that make up our thesaurus:

- **Hierarchical structure** - broader terms (BT) and narrower terms (NT), including some instances of polyhierarchy, where a narrow term can have more than one broader term.
    - BT *Curriculum*
        - NT *National curriculum*
        - NT *School curriculum*

  Every term then fits into the hierarchy by answering the question “What is this a type of?”
- **Associative structure** - related terms (RT) where a term is connected to another term, but they are not related by hierarchy.
    - *Foster carers*
        - RT *Fostering*
- **Equivalence structure** - USE and Use For (synonyms and quasi-synonyms, denoting preferred and non-preferred terms).
    - *School dinners* USE *School meals*
    - *School lunches* USE *School meals*
- **Disambiguation structure** - refinements to a term that clarify the context of use.
    - BT *School workforce*
        - NT *School workforce (academies)*
        - NT *School workforce (maintained schools)*
- **Contextual structure** - scope notes, history notes and term notes provide context and guidance for using the terms. They also give a decision making record, often useful in understanding how a term or relationship has been developed.

  References to source material are included, where possible, to demonstrate the provenance of the term and to provide additional information. GOV.UK is the preferred source for our material and provides a “master” for settling the vexed questions of capitalisation and acronyms or abbreviations.

The background to DfE’s thesaurus

The DfE and its forerunner departments, including the Employment Department, started using and developing a subject thesaurus in the mid-1980s.

The thesaurus has grown and moved with the times, and changed to meet different subject remits. As of December 2017, the master thesaurus contained nearly 8,000 terms. At the start it was used as an indexing tool in a library catalogue. Later it provided subject metadata for document and records management. More recently it was used as a tagging tool for educational Internet content.

All of these thesaurus uses were behind the scenes, and few of the department’s users were even aware of its existence.

DfE’s current taxonomy use

Everything changed in 2014. The DfE began work on implementing a customer relationship management (CRM) system to handle correspondence received by the department. It was replacing an earlier system which had used an uncontrolled, folksonomy approach to subject categorisation, where terms were devised "on the fly" to meet an immediate need and then re-used (or not) in the system.

In developing the new CRM (using Microsoft Dynamics Online), we wanted a controlled subject vocabulary for several reasons:

- to avoid unnecessary proliferation of terms
- to remove some of the guesswork that comes with a folksonomy approach
- to enable more effective browsing of terms through the use of a hierarchy
- to make retrieval more predictable, by using consistent approaches and standard formats
- to create links between the terms, teams and individuals

In Dynamics each subject term is connected to a team (or organisational unit) within the department. The act of assigning the subject team to the item determines who answers that enquiry. In this way, correspondence can be forwarded to the relevant team via Dynamics. It is very important that the subject terms are associated with the correct teams, and that the right terms are available for selection. Misallocated correspondence results in delays.

The department processes around 50,000 cases a year on the CRM system. This includes all external emails, letters, Freedom of Information requests and Parliamentary Questions. A team of trained loggers assign each case one or more subject terms from the taxonomy. The structure of the taxonomy needs to be clear and have little ambiguity, otherwise misallocation or inefficiencies can occur.

The taxonomy has another main use in the department. It forms part of our SharePoint Online deployment that runs our intranet, documents and records management and collaborative working functionality. Each team member has a staff directory entry in SharePoint Online, this is contained within the SharePoint Delve app.

We have created a set of customised fields that staff complete in their profile. Of these, two are mandatory – responsibilities and skills. Both of these fields are populated using controlled vocabularies extracted from our master vocabulary file held in MultiTes. This means that every person in the department has to interact with the taxonomy, and the agreement of our most senior managers to make this field mandatory was another major breakthrough in our use of the terms.

Ensuring that staff choose their responsibilities from the taxonomy helps us to find colleagues who can help with particular enquiries using a standard vocabulary. It provides consistency in searching, as more and more people become familiar with the terms and their uses.

Every page of intranet content is tagged with taxonomy terms, along with every article added to our News area. We will be configuring search to make the most of the subject terms to aid retrieval.

The need for pragmatism

Whilst we have followed the international standards for vocabulary construction, when it comes to practical implementation, we are constrained by the functionality of the systems we use. In an ideal world, we would use a controlled vocabulary and Boolean logic to manage search and retrieval in the way that you might in a library catalogue. However, it is not straightforward to use Boolean logic within Dynamics. Additionally SharePoint Online does not display the full hierarchy for a term when viewing a profile, even though this hierarchy can be seen during the selection process.

As a result, we need to use slightly convoluted structures to allow staff to select the most appropriate terms for their profiles.

Subject taxonomy terms are presented in Dynamics as a concatenated string with a maximum of three levels. Each “string” has to have at least two levels (a broader term and one narrower term), but some terms have a broader term and two narrower terms. This example shows both the separation of terms and the tautologous effects of the necessary refinements to cope with use in SharePoint Online:

> *Analysis : Data analysis : Data analysis (adoption)*
>
> *Analysis : Data analysis : Data analysis (fostering)*
>
> *Analysis : Data analysis : Data analysis (private fostering)*

This level of detail and disambiguation is needed to reflect terms used in both Dynamics and SharePoint. If the team is split, then the terms have to reflect both the responsibilities of the individual, and ensure that Dynamics cases are directed to the correct team.

Designing the taxonomy

From the 8,000 terms in the master vocabulary file we needed to select the core terminology to be extracted into the subject taxonomy. We started with three design principles and developed these over time through a process of engagement with correspondence loggers and policy teams. This process is ongoing. As the department's work evolves, so does the taxonomy.

Principle 1: How many levels of hierarchy would work best?
We decided to follow the rule of three, commonly used in the world of presentation and publication.

For the end users we think that three levels works well. Four would have made it more difficult to keep the hierarchy in mind and would have added an extra layer of complexity. From the perspective of being the taxonomy manager, four levels would have been beneficial. It would have given greater flexibility and helped with some particularly complex policy areas. However, there is little point in developing a system that end users might find more difficult to navigate, so three it is.

Principle 2: The level of detail must be right for each context

The master vocabulary file contains up to seven levels of hierarchy, so the next challenge was to decide how to slice the hierarchy in the best way to produce a taxonomy that gave the right level of detail for the task.

Each policy area has a different level of complexity and depth. We needed to understand the policy briefs, and determine where to pitch the level of the hierarchy to cover all of their work adequately.

We consulted the policy areas through a series of workshops using a starting point of 64 broader terms, which covered the highest levels of the department's remit. These terms were listed alphabetically on plastic sheets using a permanent marker. In each workshop, narrower terms were added and relationships between terms developed using dry wipe pens. Photographs were taken after each workshop for the record, and the annotations from each group were then erased so as not to influence the next group. The sheets could be rolled up, moved between the department's sites and used again for each workshop.

During the workshops, the policy teams used their own language to add terms to the sheets under the broader terms. At the end of the process, we translated their language into the more precise terminology used in the master vocabulary file. Then we used this information to create a subset of the master, pitched at the right level of detail. This resulted in an initial list of c700 terms, which has since developed into a taxonomy containing 2,200 terms (as at December 2017).

Principle 3: Provide multiple entry points for a browse structure.

The subject matter of a piece of correspondence is not always clear and there are often many ways of categorising it. For example, a young person caring for a parent with disabilities is both a young person and a carer. When logging correspondence about this and browsing for a subject term, a polyhierarchical approach allows for multiple entry points to the browse structure:

- BT *Carers*
    - NT *Young carers*
- BT *Young people*
    - NT *Young carers*

Sometimes this can be confusing for users who see the polyhierarchical feature as a duplication of a term, and therefore a mistake in the taxonomy. Once the purpose and benefits of this approach are explained, colleagues see the value and potential that this aspect of the taxonomy provides.

Engagement with the department

The original taxonomy development workshops were just the beginning.

We developed programmes of events to explain the taxonomy and engage with colleagues at both the general and specific level. We attend general team meetings, and various specific networks, to talk to as many people as possible about how to use the terms. These presentations vary from short, overview slots as part of a bigger meeting, to providing hour-long interactive sessions.

The hour-long workshops invite colleagues to not only find out more about the taxonomy, but to think about the hierarchical, associative and equivalence relationships in a variety of vocabularies from different settings. We show users of our vocabulary that other organisations use controlled language too, proving that it is not a “flight of fancy”, but a serious undertaking that has many benefits to our work.

Finding the “hook” that inspires a user into understanding the taxonomy, or becoming an advocate for it, is very satisfying. Staff might be intrigued by the NASA thesaurus, interested in why we have a photograph of a green sea turtle on our subject terms wiki pages, or just get carried away with using language and working out how terms connect.

It can be about setting their own work in a broader perspective, the chance to see their role in a new light, or looking at the bigger picture of what the subject taxonomy can do to streamline our processes. Whatever provokes the response, the “Ah ha!” moment is always one of the highlights of any meeting.

[If you were wondering, the turtle is there to illustrate the species taxonomy developed by Carl Linnaeus and is used as another way of encouraging discussion.]

Patterns, trends and mappings

We use the subject terms to help us look for trends and patterns, showing peaks and troughs in particular correspondence areas. This helps us to plan, and ensure that we can provide the best possible service at the right time. This might mean updating education GOV.UK content to meet particular needs at particular times, or changing a telephone queuing message to direct enquirers to alternative sources of information. The taxonomy helps to provide the management information that backs up these decisions.

A key area of mapping work is around the development of the education taxonomy that is used on GOV.UK. We worked with Government Digital Service (GDS) on the GOV.UK education taxonomy. We wanted to ensure that the mappings would "translate" between the departmental taxonomy and the GOV.UK taxonomy. There is more information available from the GDS blog posting: Finding things: how we're breaking down the silos on GOV.UK.

We knew from the beginning of this project that GOV.UK and the department would not be able to share the same taxonomy. This is mainly because of the level of detail that we need to use to map our correspondence to particular teams. However, by ensuring that both taxonomies are aligned, we get the best of both worlds because we can exchange terms and learn from each other.

Future ideas and developments

Looking to the future we are keen to develop the taxonomy into different areas including:

- using the terms to index departmental information sheets
- assigning them to telephone calls in the same way as we apply the terms to correspondence
- expanding use into the metadata associated with documents and records in SharePoint
- investigating ways that we might link terms to particular SharePoint Team Sites
- using terms in other SharePoint collaborative features such as Groups and Teams
- incorporating associative (RTs) and equivalence (non-preferred term) relationships into Dynamics to help with correspondence logging

Conclusion

At the DfE, we have designed and implemented a practical application of a subject vocabulary. We have built on the foundations and structure provided by rules and standards to put our taxonomy at the heart of what we do.

This has not always been easy, and there is more to do, but we have made a good start. We are looking forward to the next steps on the journey.

Further reading

Aitchison, J., Gilchrist, A. and Bawden, D., 2000. *Thesaurus Construction and Use: A Practical Manual*. 4th ed. London: Aslib.

Dextre Clarke, S G. (2008). The last 50 years of knowledge organization: a journey through my personal archives. *Journal of Information Science,* 34 (4), pp. 427-437.

Flett, A. and Vernau, J. (2011). Applied taxonomy frameworks. *Business Information Review,* 28 (4), pp. 226-235.

Garshol, L M. (2004). Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. *Journal of Information Science,* 30 (4), pp. 378-391.

Gilchrist, A. and Mahon, B. eds., 2004. *Information architecture: designing information environments for purpose*. London: Facet publishing.

Hedden, H., 2016. *The Accidental Taxonomist*. 2nd ed. New Jersey: Information Today, Inc.

Hedden, H. (2017). *The Accidental Taxonomist*. [online] [Accessed 08 December 2017].

International Standards Office, 2011. *ISO 25964-1 Information and documentation – Thesauri and interoperability with other vocabularies: Part 1: Thesauri for information retrieval*. Geneva: ISO.

International Standards Office, 2013. *ISO 25964-2 Information and documentation – Thesauri and interoperability with other vocabularies: Part 2: Interoperability with other vocabularies*. Geneva: ISO.