Metadata driven information systems and distributed information discovery

White Paper

Autumn 2011

tellura
the web, document and information people

Document information

This document is a white paper on metadata driven information systems and the opportunities for developing distributed information discovery services based on metadata.

Document control

Title	Metadata driven information systems and distributed information discovery
Version	0.5.20111011
Date issued	11 October 2011
Author	Dr Ian Piper, Tellura Information Services Ltd.

Table of contents

Introduction

The National Strategies were, until the end of their contract in March 2011, a set of professional development programmes for UK education professionals. They formed one of the UK Government's principal vehicles for improving the quality of learning and teaching. The programmes were underpinned by a large repository of quality-assured content:

- Core subject frameworks to improve standards of teaching and learning and maintain consistency of delivery
- Case studies of best practice
- Collections of rich media for use in the classroom.

All told the repository held in excess of 20000 documents. At the start of the project under discussion, the existing website suffered from many of the classic problems of large websites:

- Users found the website hard to use and the content hard to find.
- The site was strongly editorially focused, which meant that content was manually associated with particular locations on the site's pages.
- This in turn resulted in a delay in workflow between content being written and being visible on the site.
- It also required considerable editorial effort in keeping the site up to date.

These problems drove the organisation to consider a radically-different approach for its new content repository, central to which was the use of metadata in controlled vocabularies.

Overview of the redevelopment plan

When the site became due for redevelopment, in early 2008, the organisation decided to take a significant departure from traditional website architecture. The design of the new site was to be based on completely dynamic content discovery and navigation by use of metadata. Rather than being associated with a fixed site structure, content was free-floating: all pieces of content were stored as first-class objects in a repository. Each piece of content was to be assigned tags selected from a range of controlled vocabularies. The navigation structure of the site was also defined by a taxonomy, each navigation point (or node) of which contained a query. The query defined the content that should be retrieved when a user chose that navigation point.

For example, if a user navigated from the home page to the "Primary" landing page and then to navigation points for the subject "Mathematics" and then the topic "Assessment", she would see a collection of content that had been tagged with **Primary** from the School Phase vocabulary, **Mathematics** from the Curriculum subject vocabulary and **Assessment** from the Educational topic vocabulary.

It was immediately clear that designing a system of this type presented some interesting technical - and organisational - issues. Among them were the following:

- How would the organisation design the categorisation scheme? What controlled vocabularies would be needed to cope with the volume and breadth of content?
- How would these vocabularies be developed and maintained? What information standards needed to be used - or developed? What internal quality and workflow measures needed to be adopted to ensure good vocabulary design and a consistent approach? Finally, how would the organisation manage the development of the vocabularies as time moved on?

- How would the organisation ensure that content was effectively tagged? Clearly, if the profile of tagging was the basis of information discovery, correct tagging was of fundamental importance. This presented the technical issue of exactly how tagging would happen, and the organisational issue of how to ensure quality and consistency.
- How would navigation work? If content surfacing at specific points in the navigation structure had to meet specific criteria of metadata tagging, exactly how should this work?
- The point above hints at the final major consideration. Exactly how would the metadata, painstakingly applied to 20000 content objects, be put to effective use? Not only within the organisation, but potentially across other partner organisations, this collection of well-described content needed to be discovered, explored, shared and aggregated to make the whole process worthwhile.

Designing the metadata eco-system

As a fundamental component of the content creation and discovery processes, the design and development of the environment for managing metadata proceeded in a manner highly integrated into the site development process. Early decisions then included

- the format of vocabularies to be used,
- the identification (and, where necessary, development) of any existing vocabularies that would be applicable,
- the scope of vocabularies that would be required and
- processes for making vocabularies available to the content.

As discussed towards the end of this document, the latter decision and its consequences led the design team towards some opportunities for future use of vocabularies in federated content discovery.

Taxonomy file formats - Zthes and SKOS

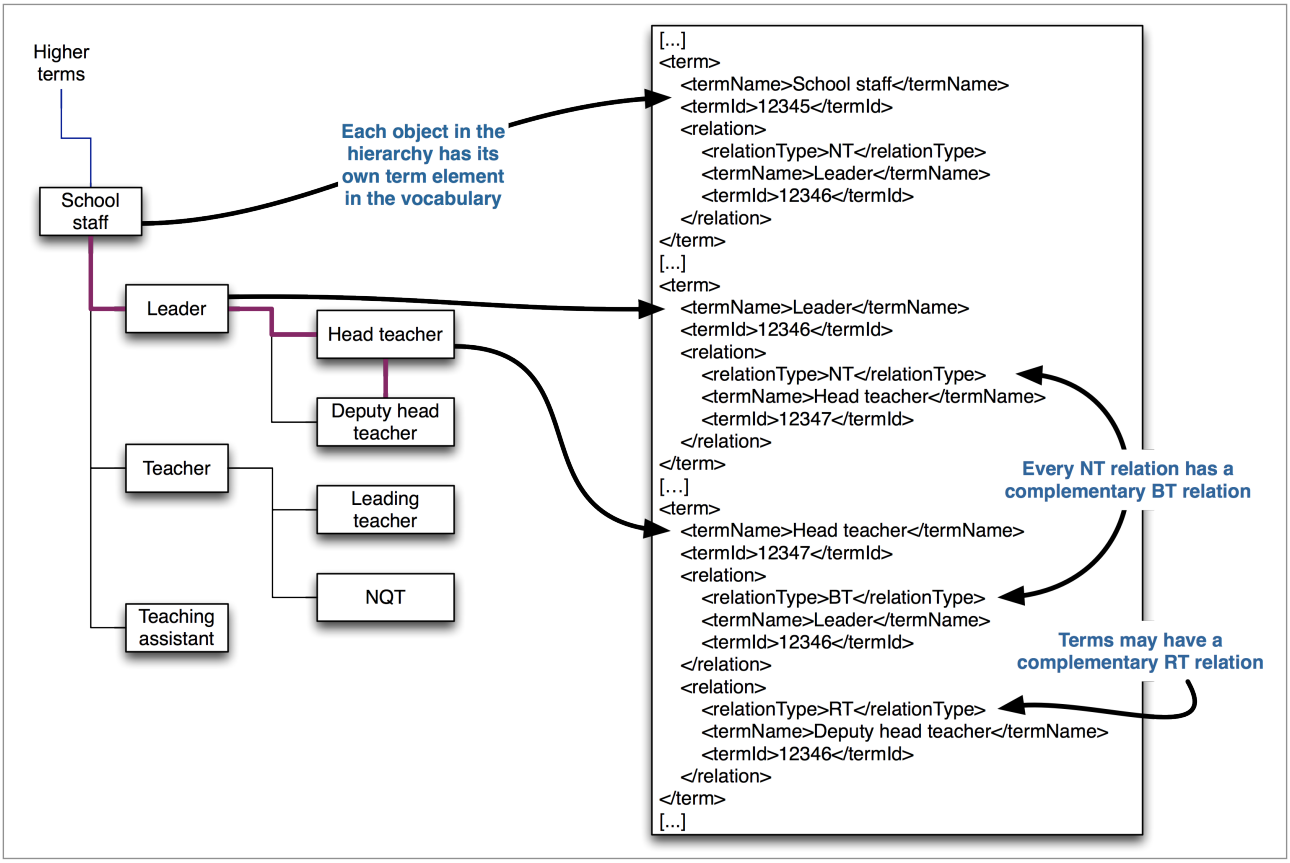
An early design decision was to use Zthes as a vocabulary format. The team selected Zthes for a number of reasons:

- It was a commonly-used standard
- It had reasonable tool support
- Unlike a hierarchical xml-based format, it allows storage of multiple types of relations between terms within and across vocabularies.
- It has a simple mechanism for extending the structure where it is necessary to store additional metadata (using the termNote element and its attributes).

Zthes is an xml-based format, but unlike many xml formats it does not manage hierarchical information structures by having a hierarchical structure itself. Instead, a Zthes file is made up of essentially flat elements (called terms). The tree structure of the taxonomy is implied, by having elements (called relations) that store relationships between terms. The figure below shows a schematic representation of part of a Zthes file, showing how the hierarchical elements map onto the taxonomy structure via broader and narrower term elements (BT, NT and RT respectively). This is simplified for the purposes of illustration.

As the figure shows, each object in the hierarchy has a <term> element in the taxonomy. These are all presented as top-level terms (that is, lower-level terms are not contained within higher-level terms). The hierarchical structure is stored by means of BT and NT relations. So Leader is a narrower term of School staff, and School staff is a broader term of Leader. Also, Head teacher and Deputy head

teacher are not in a NT-BT hierarchy but rather a RT-RT relationship (because while the concept of deputy head teacher is *related* to that of head teacher, it is not a *narrower* term).



During the course of the project the team converted the vocabularies from Zthes to SKOS (Simple Knowledge Organisation System) format. This reflects both a development in the needs of the organisation and the enhanced capabilities offered by an RDF-based format for semantic web application development. It is worth noting in passing that the team became well-versed in processes for inter-converting vocabulary file formats, as there were a large number of such formats in use!

Given the flattened nature of the structure it was not feasible to develop Zthes vocabularies using conventional xml editors. Instead the team identified a commercial vocabulary editor which provided an adequate solution for vocabulary development.

Another component of the metadata eco-system was the scope of vocabularies to be used. The idea of a single taxonomy was quickly dismissed: the likely collection of vocabularies in the system had a broad range of focuses, and it is a general rule of taxonomy building that aggregations of concepts should be on the basis of broad similarities. The decision here was to build a range of topic-focused vocabularies, each as a separate taxonomy. This decision had a number of important consequences:

- It was possible to have strict aggregation of concepts into vocabularies that really all were about the same thing
- It became easier, later on, to build profiles of tagging for content objects (i.e. one could describe a content object as being tagged with concept A from the Phase vocabulary, concept B from the Role vocabulary and so on - the vocabulary name itself became part of the tagging profile)

- Vocabulary management became easier, in that several vocabularies could be worked on separately and simultaneously.

Another factor in this decision was the availability of existing vocabularies. An earlier government education project had produced various vocabularies (mostly in Zthes format) and it made good sense to incorporate these into the suite of vocabularies used in the project. In some cases, these needed a lot of work to update, as concepts had changed in use or become deprecated in the six years or so since last being used. This highlighted to the team the importance not only of creating useful vocabularies but also of keeping them up to date, and pointed to the benefits of managing vocabularies in a shared and open environment. I will return to this topic later in the case study.

Finally, there needed to be a way of getting vocabularies into the same environment as the content and then associating content objects with terms from various vocabularies. Without going into technical details, the team decided to build on Drupal's existing taxonomy infrastructure, extending this out to allow many interconnected vocabularies. The project also used a custom development of a commercial tagging tool. This allowed content developers to check out a content object, allocated tags to it and check it back in. While this essentially duplicated Drupal functionality, it added features such as business rules. One of the lessons of the project was that in fact such business rules were better managed by organisational process, and the experience pointed towards an alternative approach to the tagging of content. See the section on future ideas near the end of this document.

Building and management of controlled vocabularies

To some extent the design and development of the vocabularies was an organic, ad-hoc process. From early discussions with educational specialists and the Department the designers knew that there were some aspects of information categorisation that they would need to include, such as:

- Stages of education – Phase (Early Years, Primary and Secondary), Key Stage, School Year
- Curriculum subject and subject specifiers
- Role or Audience

A very important source of taxonomy terms turned out to be the subject experts who tagged the content. As they worked through the tagging process they would often come up with concepts that were covered in the content but not captured in a tagging vocabulary. The team collected these concepts together and periodically reviewed where they belonged. Often this resulted in a new vocabulary, perhaps in response to a large amount of complex content that had been created in a specific topic area.

Other vocabularies were created in the expectation of being used, but ultimately were hardly used at all. Examples included the Language and English region vocabularies. Very little of the content ended up in languages other than English, and very little had any regional focus.

Finally, it pays to expect the unexpected. One of the most useful vocabularies, Educational topic, came about as a result of the need to have a general "dumping ground" for terms that had no obvious home. Later in this section one of the house rules describes how vocabulary terms should be kept in topic-focused vocabularies that represent those concepts. As the project developed a number of vocabulary terms were created that could not easily be placed within any existing taxonomy. So the team created a new vocabulary called Educational purpose and these orphaned terms were placed in it. It became clear as time went on that in fact there were clear themes developing in this vocabulary, and the team restructured the vocabulary along these thematic lines. The vocabulary was renamed Educational topic and became one of the most widely used and useful vocabularies.

Developing a set of house rules

As the vocabularies developed in number and complexity it became clear that there was a risk of confusion unless some house rules were adopted. The rules used were fairly straightforward:

- All terms across all vocabularies should have unique identifiers. The team developed a tool for generating MD5 hashes of a randomly-generated string, and the resulting 32 character string was used, as part of a URI, as the identifier.
- All terms in a vocabulary should be specifically about the topic of that vocabulary. For example, the School Year vocabulary contains only terms for school years, and not for Key Stages or phases.
- If terms are arranged in a hierarchy they should reflect a genuine hierarchy. For example, Deputy headteacher is not a narrower term of Headteacher (though it is a related term (RT), but Leading teacher is a narrower term of Teacher.
- Where a term already exists in a different vocabulary it is used as a relationship rather than creating a new term with the same name.
- As a consequence of the above point, different terms should have distinct names for clarity.
- Term names are always singular except where the sense of the concept was essentially plural. So the term Teacher is used rather than Teachers. However, National Occupational Standards (NOS) is plural because the term itself is always used in that way.
- Sentence case capitalisation was used throughout except for proper nouns and government initiatives. So the team used Special educational needs (SEN) (for a phrase in common usage) but Disability Equality Scheme (DES) (for an initiative).
- Where acronyms are used, they should be preceded by the expanded term. For example, Communication, Language and Literacy Development (CLLD).

Naturally, house rules should be driven by the need of the work rather than dogma, as it is more likely to be maintained by the users if they agree with the reasons for adopting the rules. The exception above is the use of singular terms: the team could just as easily have standardised on plural terms, but the important thing was to have a consistent approach.

Tagging content objects and building a navigation scheme

As mentioned earlier, the fundamental design principle for the new content repository was that of information in a flat repository being dynamically surfaced at navigation points by virtue of tags applied to the individual pieces of content. This required that the content objects should be properly tagged, and also that the navigation points in the site hierarchy should have queries that are designed to retrieve content based on its tagging. For the latter purpose the team adopted the contextual query language (CQL). CQL allowed a content query to be expressed in something close to plain English. To allow this query to do something useful, the team also developed extensions for Drupal to allow content to be identified from such queries. Essentially this involved a conversion of a cql query into a SQL query, but this document is not going to go into further detail.

The site development proceeded with content tagging and navigation development happening simultaneously. This had both advantages and disadvantages over the alternative approach of tagging first and then building the navigation. The advantages were principally in the area of accelerated development: the timescales demanded parallel development, but it meant that the tagging and navigation teams worked very closely and productively together. The principal problems arose from the fact that this hadn't been done before, so all of the team was learning. One example was that of over-tagging. The temptation when tagging content was always to put in every possible relevant metadata tag, in order to ensure that the content was being adequately described. However, this had the consequence that items often came up in many places, some of them unexpected. The complementary problem was over-complex queries. The early queries built tended to include all possible descriptors for a particular navigation point, with the result that sometimes no content was returned. The tools available also made it difficult to check the validity of queries, which also resulted in navigation points with no content. The CQL query language allowed for the development of complex Boolean queries using AND, OR and NOT clauses. These proved to be very useful in

tailoring queries, and the business process that evolved - making changes to the query, checking, making changes to the tagging, checking - allowed fine control over the amount and nature of content returned at a navigation point.

The process of refining both tagging and query building was necessarily iterative, and the team developed some rules of thumb over the course of the project:

- Use the minimum number of tags to adequately describe the content. If in doubt, don't use a tag
- Avoid the use of "OR" clauses in queries. These expand the number of results returned, sometimes dramatically.
- Avoid the use of "NOT" queries completely.
- Avoid using more than four clauses in a query.
- Don't use more than one clause from a hierarchy of possible clauses taken from a vocabulary.
- Choose the lowest level of tag available from a vocabulary, and don't choose its parent tags in addition. So if a piece of content is tagged with Leading teacher it should not also be tagged with Teacher.

Putting it into practice

In many respects this was a bold project, in creating a model for information discovery that had not previously been widely explored. Not only did it require some sophisticated software development, it also required a high degree of collaboration between disciplines: taxonomists, subject experts, education practitioners, content managers, editors, writers, not to mention information architects, user experience specialists, graphic designers and, of course, software developers.

In practical terms the coordination of these disparate disciplines worked extraordinarily well, due in part to the Agile project methodology adopted. The iterative nature of work was also very helpful, in establishing the value of re-visiting tagging, query-building and even content editing where necessary to enhance the value of the content.

As indicated in the section on tagging and navigation building, not everything worked immediately. Early over-tagging resulted in insufficient discrimination in content retrieval, while early query design sometimes gave unexpected results. Iterating over the work ensured that it was continually being refined, and within a matter of months the content delivery was more or less optimum. More to the point, the system was now flexible enough to easily cope with new content. On the whole, this content appeared where it was intended to appear, and did so immediately.

The project was well-received by its various audiences, and usage figures increased dramatically over the earlier site. The success of the design in ensuring that content would correctly surface in the most appropriate locations within the site could be largely ascribed to the innovative use of metadata vocabularies in tagging the content.

One aspect of the system that has not been mentioned here gave an additional benefit of the use of vocabularies. The team adopted the Solr faceted search tool during the course of the project, and this allowed for the enhancement of content searching by providing a secondary navigation based on facets. This is not the right place for a technical review of Solr, but in summary the tool provides a full-text searching feature supplemented with a filtering mechanism. Results from a Solr search can be presented in such a way as to allow them to be refined based on the metadata terms (called facets) used to tag the content in the resultset.

The end of the project

The new government elected in the summer of 2010 decided to bring the project to an end, in common with many other public sector projects. The team had a number of developments still in planning at the time the site closed, and these remain as opportunities. The rest of this document

discusses those opportunities pertaining to the use of metadata. In particular, I will be covering the Open Vocabularies Service.

Background to the Open Vocabularies Service

Over the course of the project the team began to research the design of better taxonomy development tools, particularly with the aim of having a completely web-based experience and a distributed model of tagging. This work culminated in the development of the Open Vocabularies Service (http://openvocabs.org/), a web-based repository and toolset for the creation and management of controlled vocabularies. This tool was never implemented in the main site, due in part to the closure of the project, but it was spun off later as a separate open source product.

The genesis of the development came from a number of lessons learned within the National Strategies project.

First and most important, the team became aware over the course of the project of the potential value of federated information discovery. There were a number of partner organisations in the education sector, and broadly speaking all of these took a similar approach to delivery of content on the web. That is, for the most part they were isolated silos of content - local data stores, local taxonomies developed with tactical rather than strategic aims and no real plan to share or link content with other organisations.

Given the observable trend represented by the growth of the Open and Linked Data movement, the team began to consider whether the National Strategies content repository could take on some more open aspects, to make it easier both to make use of content from other repositories and to syndicate its own content out to other sites.

The team began to consider how such ideas might be developed into concrete features within the website. There were some clear and relatively straightforward possibilities:

- The current site already had the capability to publish any piece of content as RSS (this is core Drupal functionality). It would be only an incremental piece of work to publish content as RDF (as some organisations such as the BBC are already doing: see, for example, this page: http://www.bbc.co.uk/nature/species/lion and this RDF version: http://www.bbc.co.uk/nature/species/lion.rdf).
- Each piece of content already had a unique id, internally determined by Drupal. It would be straightforward to enhance the service and allocate a URI to each piece of content. This, with the RDF feed, would allow other organisations to tap into the content directly.
- Content could be marked up with microformats or RDFa tags to enhance the value of the content to discovery services.

However, there appeared to be other areas in which novel functionality could be created. Even with the above features, the core innovation of the NS site - its use of metadata - was still used within a silo. The vocabularies could be exported and given to other partners, but that in itself is not a solution to the problem of shared metadata. The team began to examine the feasibility of separating out the metadata from the content - having the content tagged against a remote, open (and therefore potentially shareable) metadata repository.

If it were possible to tag content against a remote metadata repository, using a URI mechanism to identify the metadata terms within the controlled vocabularies, then in principle other partner organisations could do likewise. This in turn gives rise to an intriguing possibility: if a variety of organisations choose to tag their content against the same shared metadata repository, then in principle there is a mechanism for distributed information discovery based on that shared metadata. There would be a number of technological hurdles, of course:

- Is it possible to develop an abstracted mechanism for tagging content that would avoid having partner organisations having to customise their content management systems?

- How could the shared metadata repository be made aware of the existence of all of the services that make use of its concepts (in other words, exactly how would distributed information discovery work?).

Several ideas were developed to address both of these potential hurdles. This is not the right place to go into detail, but for the first question the team developed a number of browser plug-ins that would directly tag Drupal content on demand. For the second question, the team looked at the innovative use of JavaScript and HTML markup used in tools such as Clicktale, and concluded that similar markup might be used to create a two-way conversation between the metadata repository and its various subscribing websites.

Given the ideas developed so far, the first step seemed to be to develop the appropriate metadata tools and repository. The development had a number of core design principles:

- It would provide a repository for creating, importing, exporting and management of hierarchical vocabularies.
- It would provide web-based tools for all vocabulary management processes.
- Every term in every vocabulary would have a guaranteed unique identifier.
- Every term identifier would be a URI.
- It would have a compelling user interface and as low as possible a requirement for help or training.
- It should be responsive and dynamic in its behaviour.
- It should provide tools for tagging content within remote content repositories

Description of the core product features

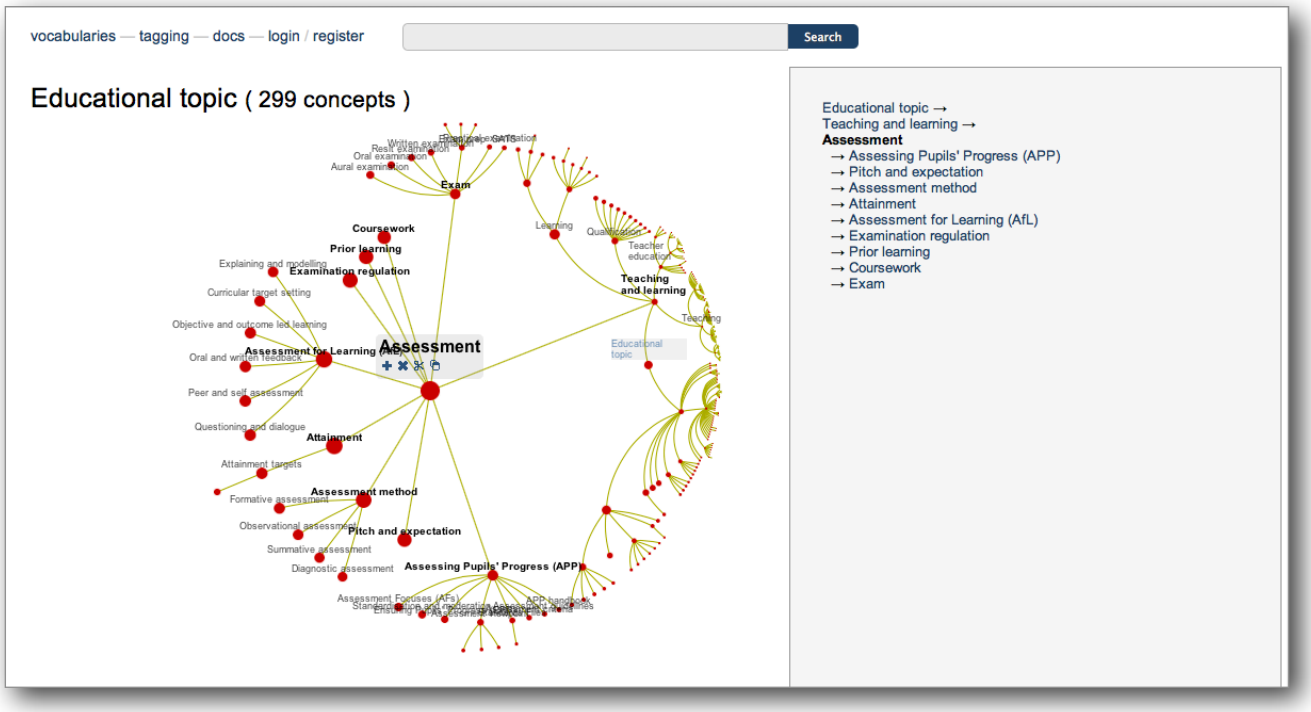
The ensuing development project had as its aim the creation of a proof of concept for as many as possible of the ideas above. Here is a summary of the services created to date. These can be seen at http://openvocabs.org/.

The repository

The service consists of a web application running Python/Django and running over a MySQL database. The use of Django, as a MVC pattern design, allows an abstract object-oriented data model for metadata vocabularies to be stored in a concrete database form. All of the individual terms across all of the vocabularies have enforced uniqueness, and the data model allows for a simple definition of hierarchies. These hierarchies are exportable in RDF format (strictly speaking, they are SKOS files - Simple Knowledge Organisation System). The repository can also be populated by importing a suitably formatted MS Excel spreadsheet - useful, since many organisations seem to use Excel to model taxonomies.

The editor

The services provides an interactive browser and editor, written using an open source JavaScript library (the JIT library). This can be seen in the figure below. The arrangement of terms is called a hypertree. Clicking on any visible node will cause the display to move to centre that node and rearrange the hypertree. Using this it is possible to navigate in real time up and down the hierarchy of any vocabulary (in the figure below, I have navigated down to the second level of the Educational topic vocabulary: Educational topic > Teaching and learning > Assessment).



If the user is subscribed and logged in, he can create a vocabulary, edit terms in existing vocabularies and delete unwanted vocabularies. Editing includes creation of new terms, cutting and pasting individual terms or whole trees of terms and deletion of unwanted terms.

Simple search tools, provide auto-complete functionality to allow for simple identification of existing terms with a given name. Template tools allow for a vocabulary to be viewed in different formats, such as a web page navigation view. This allows people who are creating navigation vocabularies like that used in the NS website to test how a particular structure will look in a real website.

Tagging tool

The service provides a simple tagging tool. This is a plug-in for the Chrome browser. It is built using HTML, JavaScript and CSS. When used in conjunction with a Drupal website, it works as follows. The user who wishes to tag a piece of content logs in and navigates to the content object. He then selects the Edit tab for the content object. At this point the plug-in is activated and provides a pop-up window containing a list of vocabularies. Clicking on any of these opens it up to show its constituent concepts. Choosing one of those adds it to the list of terms that are tags for the content item. An auto-complete search box enables the user to rapidly see whether there is a concept with a given name (type in "c-h-e-m" and all terms containing "chem" anywhere will appear. Again, the editor can simply add those terms to the tag list. Clicking the Save button immediately saves the content object with its new tags.

This plug-in is a very rough proof-of-concept (as indeed is the whole Open Vocabularies Service at present!) but it would be a fairly straightforward process to elaborate similar plug-ins for other browsers and other content management systems.

Opportunities for further development

It is worth repeating - this is a proof of concept. It is buggy, has a poor user interface, has limited user-based permissions and has some limitations that restrict the size of vocabularies that can display in the hypertree view. None of these detract from the core proof of the concept - it is a real shareable metadata repository, with real potential for future development.

So what are the areas in which development is needed? Some, in no particular order, are:

- The known bugs need to be fixed
- The user interface needs a lot of work
- The hypertree view should be supplemented with other views such as a simple tree with disclosure points
- The performance and display issues should be fixed so as to allow display and management of any size of vocabulary
- The permissions should be improved so that vocabularies are protected from accidental deletion of terms (or indeed of whole vocabularies)
- The tagging tool needs to be refined and elaborated to work with other browsers and other CMS products

The final area for development is more speculative and longer-term. It is the area alluded to above, in which a future implementation of the Open Vocabularies Service would know what services were using it to tag their content. This more semantically-aware service would form the basis for an information discovery mechanism based not only (maybe not even) on full text but on the tags allocated to content. For example, a user of such a discovery service may want to find content in any participating information service that is aimed at head teachers. Knowing that those services will have used common vocabularies for tagging their content increases dramatically the confidence that the user can have in the quality of the discovery process.

I have described this as speculative because it requires a number of potentially complex developments: developing a mechanism for storing tag URIs within a page, for example, and building an efficient mechanism for having a registry of tagged content. Initial experiments and knowledge of modern services such as Clicktale indicate that this is a perfectly achievable aim.

Dr Ian M Piper
Tellura Information Services
http://www.tellura.co.uk
ian.piper@tellura.co.uk
Autumn 2011