The Accidental Taxonomist: An interview with Heather Hedden

Interviewed by John Horodyski, Manager, Digital Programming, CBC

Received (in revised form): 19th March, 2012

Heather Hedden

is a taxonomy consultant with Project Performance Corporation and author of the book 'The Accidental Taxonomist'.[1] Previously she was the taxonomy manager at First Wind Energy LLC, and prior to that an independent consultant offering taxonomy development, training, and indexing services through Hedden Information Management. Heather also teaches online workshops in taxonomy creation through the continuing education programme of Simmons College Graduate School of Library and Information Science and has given numerous conference presentations and workshops on taxonomy topics.

98 East Riding Drive, Carlisle, MA 01741, USA
Tel: +1 978 467-5195; E-mail: heather@hedden.net

Abstract In this interview for the Journal, Heather Hedden discusses with John Horodyski some of the interesting challenges with taxonomy based projects, innovations in taxonomy software solutions, as well as her thoughts on taxonomy best practices in design.

KEYWORDS: taxonomy, controlled vocabulary, metadata, indexing, classification

INTERVIEW

JH: Heather, would you tell us a little about your professional background?

HH: I consider myself a taxonomist, which is still not a widely recognised profession, but it is a recognised and valued skill set and, increasingly, a job title. The skills are analytical, involving initially analysing a set of content and its intended uses and audiences to develop a data model for metadata, then developing controlled vocabulary that balances specificity and inclusiveness, including synonyms as needed, and involving hierarchical and other relationships between concepts.

I entered the field first as an indexer, using controlled vocabularies to index periodical articles for what was then Information Access Company (later Gale, and now Cengage Learning), a database vendor to libraries. After a few years indexing, I took a position as a controlled vocabulary editor in the group that developed and maintained the taxonomies. We also mapped our taxonomies to those of third-party vendors. While in that position, we had a multi-year project of revamping all the controlled vocabularies to put them into correct hierarchical structure, which gave me great experience in working with taxonomy standards.

I later worked as a taxonomist at Viziant Corporation, a start-up developer of enterprise search software. Through this position, I learned the similarities and differences in creating taxonomies for human versus automated indexing. Then as the taxonomy manager at a wind energy company, First Wind, I developed taxonomies both for search and for content classification in SharePoint. Now

in consulting I can apply my full-range of taxonomy skills to various challenging projects. I started my own business, Hedden Information Management, in 2004 and in January 2011 joined Project Performance Corporation.

JH: How would you define taxonomy?

HH: Taxonomies are structured sets of terms used to tag or index content assets, and also used for the retrieval of that content, whether through browsing or searching. Another name for a taxonomy is a controlled vocabulary. Typically, the taxonomy for a content repository comprises multiple sets of terms grouped by type, such as Topics, Places, People, Activities, etc. Terms within each set/type could be a short list of a dozen terms or so, or hundreds of terms arranged in a hierarchy of broader and narrower terms several levels deep.

JH: Can you, in some detail, describe the types of project that clients ask you to address most frequently? What are the common pitfalls?

HH: The types of project that I have worked on, both through Project Performance and Hedden Information Management, have been quite varied and have included retail website taxonomies, public informational website taxonomies, and internal content management system taxonomies. Industries have included publishing/advertising, health information, financial services, retail, and non-governmental organisations. I don't specialise in any particular type of taxonomy, so there has been no pattern to the types of project.

Increasingly taxonomy projects are no longer just designing a new taxonomy from scratch, but rather taking various existing taxonomies, controlled vocabularies, navigation schemes, classification codes etc, and integrating and updating/revising them. There could be multiple legacy databases, each with their own categorisation scheme, and they need to be merged. When dealing with legacy-controlled vocabularies, the challenge is always deciding how much of the old system of classification should be retained and how much it should be changed. Users may be familiar and comfortable with the existing categories and topics, so they shouldn't be completely changed without good reason.

Although a taxonomy can serve many needs (manual tagging, automated classification, end-user browsing, end-user search, etc), the primary need is an important factor in how to design the taxonomy. Sometimes we encounter the problem of a client not being sure what its primary use for the taxonomy will be.

JH: What role do taxonomies play in implementing business strategies?

HH: While all taxonomies play a role in business strategies, what we call an enterprise taxonomy or business taxonomy addresses strategy more directly. Taxonomies enable employees quickly and accurately to access desired content assets that are crucial to an organisation's business. The structure of the taxonomy, specifically the designation of specific hierarchies or facets, needs to match the users' needs. For example, if geographic location is important to know, then it needs to be a facet in the taxonomy. But if it's not, it shouldn't be included. Trying to cover all the bases by designing a taxonomy that may be more complex than necessary can have a detrimental effect, whereby users may avoid using the taxonomy altogether. This is especially the case for the editorial users who assign metadata/tag content. If they have too many fields to tag, they may just ignore some or use a few common tags instead of spending the time to tag accurately.

JH: Hierarchy vs Facet … is it really a debate at all?

HH: Yes and no. When designing an overall taxonomy, it's true we have two primary models in mind to choose from: hierarchical or faceted. A hierarchical taxonomy involves having each term in a hierarchical relationship with another term, where relationships are the type broader term/narrower term, aka parent/child. For example, a hierarchy of terms could be: Business > Marketing & Advertising > Direct Marketing > Database Marketing. A faceted taxonomy, on the other hand, has several (usually three to seven) attribute categories of terms, which are used in combination to describe a particular content asset. For example, facets for marketing could be: by Age Group, by Ethnic Group, by Industry, and by Region.

Hierarchies and facets are not mutually exclusive however, and can be combined within the same taxonomy. It is possible to have a hierarchy within a facet, such as a facet for something broad, such as 'Topics'. It is also possible to have a taxonomy that is accessed first hierarchically, and then, once the narrowest term is selected, facet choices can become available. This latter scenario is common in ecommerce product taxonomies, whereby the user starts off navigating through a series of product categories from broader to more specific. At the most specific product category level, facets become available allowing users to narrow the selection further — for example, by price, brand, and other features.

The choice of design approach depends on the content. In product taxonomies in particular, it often makes sense to have a hierarchical taxonomy at the higher levels, and then facets at the narrowest categories, which share many common features. A taxonomy for content management or other digital assets is more likely to start with facets and perhaps have a hierarchy within one or more of the facets, such as Topics. Sometimes however, it is difficult to determine which structure should be the starting point, and then you may have debate.

JH: Turning to taxonomy and technology, there are some interesting taxonomy-based software solutions for end users in the marketplace. How do you find the right balance between the intellectual activity of taxonomy design and technological solutions?

HH: Creating a taxonomy is an intellectual process, whereas implementing for utilisation requires technology. Additionally, there are tools that support the taxonomy design and creation process.

Whether it's software for taxonomy design and management or software for search and retrieval by means of a taxonomy, it is usually a form of database management system. From a software standpoint, a taxonomy is just content (even though it is completely different from the content it is designed to retrieve). Taxonomy management software is, of course, not end-user software; it is software for taxonomists. But if you research 'taxonomy software' that is what you end up finding. Examples include Synaptica, Data Harmony Thesaurus Master, Smartlogic Semaphore Ontology Manager, Mondeca ITM, and MultiTes.

The software a content searcher uses is usually a kind of content management system (CMS), SharePoint, or a web-enabled database. Manually indexed content is typically stored in a database management system, with content or a link to content in one field, and taxonomy terms in another field.

Autoclassification software, on the other hand, makes use of algorithms or other technologies to match content files to taxonomy terms automatically without a human indexer, and such indexing is done

'on the fly' at the time of search, so a database system is not used. Examples of autoclassification software that also include taxonomy management capabilities are Data Harmony MAIStro, Smartlogic Semaphore Classification Servicer, ConceptSearching, Mondeca CA Manager, Nstein Text Mining Engine, and SAS Enterprise Content Categorization. Most autoclassification systems however, work only on text-based documents, not image or other media files.

While indexing can be manual (by human indexer intellect) or automated, taxonomy creation always requires human intellectual input, even if software technology may suggest taxonomy terms. Technology supports the development and use of taxonomy but does not replace it.

JH: Tell me more about your teaching experience at Simmons College? How has that experience affected you?

HH: I've been teaching online continuing education workshops through Simmons College Graduate School of Library & Information Science since spring 2006. I began with teaching a four-week online workshop 'Creating Website Indexes,' and in 2008 I added my second, five-week online workshop 'Taxonomies & Controlled Vocabularies'. I stopped offering 'Creating Website Indexes' in spring 2011, as web technologies have evolved. 'Taxonomies & Controlled Vocabularies', which is platform-independent, has remained popular. I typically teach it four to six times per year, and it fills up a couple of months in advance.

While I learned some additional things about taxonomy from creating and teaching the course, mostly I learned how others, as represented by my students, view taxonomy. This has helped me be a better consultant in dealing with my clients. I've also come to appreciate the diverse background of people that have an interest or need to create taxonomies. Finally, it is worth noting that my online course was the source of core material for my book, The Accidental Taxonomist. I wouldn't have written the book without the course as a starting point.

JH: What other techniques have you found that have really helped companies understand taxonomies and their business ecosystem, not just top-down but bottom-up throughout an organisation?

HH: To help a company understand taxonomies and how it fits into their system, consultants will often start a project with a customised on-site workshop with key stakeholders. The workshop both responds to a particular situation while also presenting general taxonomy principles and best practices. The workshop is both top-down, by looking at general organising principle, and it's bottom up, by looking at specific use-case examples.

JH: Who tend to be the key stakeholders in the end-user companies?

HH: Stakeholders include the taxonomy project owner or manager, someone from higher management who might be the taxonomy 'sponsor', existing or expected taxonomy editors, senior indexers or indexing managers if any, an IT or web person responsible for technical implementation, an information architect or someone responsible for user-interface design, and either multiple representative users of the system from different departments if internal, or a customer service representative if the system serves external users.

JH: If people want to find stuff using a particular kind of mental map and they can't relate to the user-interface and the taxonomy,

then basically you have a broken system. They'll obviously revert back to sorting through CDs, back issues or pestering a lot of people. Do you agree?

HH: Yes, occasionally some people will revert to the old ways to look for information, but as long as it's just a matter of searching, and not some other process that involves tracking in the system, it does not mean the system is necessarily broken.

To facilitate adoption of the new system, not only is education and training required, but it's also important to engage the users in the design process for the taxonomy and user-interface, so that it's responsive to their needs. This serves as a means of obtaining buy-in and also training the users in the new, yet-to-be-built system. Since it is likely that not all users can be involved in the design phase, those who are included should be selected carefully to include those who have influence over other users.

JH: Do you see any trends or best practices in taxonomy design?

HH: It's difficult to discern taxonomy design trends, at least in internal/enterprise taxonomies, because each taxonomy is unique to the needs of the organisation. I will say, though, that as taxonomies become better understood and their benefits become more appreciated, there is an increasing openness to accept more features, such as the combination at different levels of both hierarchical and faceted navigation; the use of both variants/synonyms and hierarchies for navigation; and the addition of term notes (scope notes) and other attributes for terms.

We do see a trend in internally-used taxonomies, that multiple legacy taxonomies from different departments or for different product/service lines are being merged or combined to create enterprise-wide taxonomies. As taxonomies become more common, more taxonomy work centres around revising, merging, and sometimes even translating taxonomies, rather than creating them from scratch.

As for best practices, these have not changed much in recent years. The best practices for taxonomy design are based on the ANSI/NISO Z.39.19 'standard', entitled: 'Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies'.[2] It also corresponds with ISO 25964-1: 'Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval'[3] (also published as the British Standard BS ISO 25964-1). These are 'guidelines' after all, so there is flexibility in their application, and new trends may evolve.

I write taxonomy management guidelines for my clients near the conclusion of a taxonomy development project, and these guidelines take into consideration these industry standard guidelines, but also provide more specific guidance for particular usage, style, and implementation of a client's own unique taxonomy. A taxonomy will continually evolve and grow, and I won't always be there for every past client. These guidelines have helped clients manage their taxonomies on their own in accordance with best practices.

References

1. Hedden, H. (2010) 'The Accidental Taxonomist', Information Today Inc., Medford, NJ.
2. National Information Standards Organization (2010) 'Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies', ANSI/NISO Z.39.19, NISO Press, Bethesda, MD.
3. International Organization for Standardization (2011) ISO 25964-1: 2011 'Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval', ISO, Geneva.