Book Review: The Accidental Taxonomist

Reviewed by Michael Upshall

Hedden, H., 2010. *The accidental taxonomist*, Medford, N.J.: Information Today.

Heather Hedden has written an excellent introductory manual for anyone involved in setting up, running or expanding a taxonomy or thesaurus. Unlike many books on the subject, this is one for the practitioner, based on lots of practical experience — as Patrick Lambe describes it in his foreword, "this is taxonomy from 100 feet".

Hedden very helpfully does not get too doctrinaire about the distinction between taxonomies, thesauri, and other classification systems, nor about the many and varied capabilities of taxonomy software. This is helpful, because in most cases nowadays a combined approach is used to ensure the classification adopted has the greatest possible value. It turns out when looking in detail at taxonomy software, for example, that many of the products available have modules that achieve similar goals, although of course they might all be labelled differently.

Hedden is also sufficiently aware of what is going on in the world to have looked at taxonomies in a wider context. She explains, for example, the difference between the term "thesaurus" in Roget's Thesaurus, and that used today often for machine-based indexing and searching.

One of the difficulties Hedden faces, paradoxical for a subject area that is based around classification, is that there is so little agreement on terms. It is not too difficult to define what a taxonomy is or does; this is handled in the first chapter, identifying three functions:

1. Indexing support
2. Retrieval support
3. Organisation and navigation support

But Hedden hits some problems in the following chapter, Creating Terms, since the term "concept" is described as being "any of node, object, individual, entity, instance, cluster, wordset or taxon" – an indication of the confusing lack of agreed terminology in this area. With this many terms it is not surprising that taxonomy is seen as a forbidding subject. Hedden sensibly opts for a single term within her book, and where there is a standard around, such as Z39.19, she refers to it.

Nonetheless, once a term and its usage has been agreed, you find later in the book different

interpretations and usages start to creep in. For example, we are told that Z39.19 mandates the use of lowercase for terms in a thesaurus ("apples", not "Apples"), and no inversion ("commercial loans", not "loans, commercial"), but in chapter 6, where she discusses indexing by hand, she recommends adding plenty of phrase inversions to facilitate retrieval. The problem is not that Hedden is inconsistent, but that this is an inconsistent body of practice with very little standardisation. Effective use of taxonomy often seems to be to use multiple methods, which may not always be consistent with each other.

One reason for the lack of standardisation is because the field described in the book is so vast. *The Accidental Taxonomist* includes creating the navigation for a website, which might comprise only a handful of terms, but also examines the compilation of monster subject-domain thesauri such as Inspec, which will have hundreds of thousands of terms in a tightly controlled hierarchy, as well as upstarts such as faceted searching, which has become the indexing of choice for many e-commerce websites. The methodology for creating and managing taxonomies for each of these three areas will vary widely.

I suspect the book will be valued for two things in particular: first, it provides practical guidance from someone who seems to have turned her hand to most aspects of taxonomy creation. For each process, there is a recommended procedure, which users can adopt or not as they choose, but which nonetheless provides a considered framework the reader can use in developing a plan.

The other benefit is her very clear and objective comparison of human-based and machine-based indexing. Advocates of each system seem to spend their lives complaining about the other, but the truth is as so often happens somewhere between the two. Chapters six and seven compare manual and automatic methods of indexing and cover the many and varied tools now provided by taxonomy software companies, including (but not limited to) automatic categorization, entity extraction, applying business rules, even, with some software tools, automated taxonomy generation.

The book also mentions ontologies, although it would be fair to say that the book is not really a guide to what ontologies do. This is unfortunate, since ontologies are transforming the way that content is indexed and retrieved, and they have introduced such a major change in the way that taxonomies are created and managed that (as they say) things may never quite be the same again. It may seem simply a case of changing the label from "taxonomy" to "ontology", but there are some fundamental differences between the two, and this "creative disruption" has already transformed the taxonomy software market.

Overall, it is difficult to imagine anyone, even an experienced taxonomist, not finding something of value in the book. This is well worth reading, and in fact Heather Hedden's blog (http://www.hedden-information.com/blog.htm) continues this impartial and informed tone admirably to cover the three years since the book first appeared.