Coming to Terms with Keywords

The following is a summary of ideas presented at *User Focus*,
the UPA DC usability conference, October 12, 2007.

This information is also available (and updated) online at: www.ipgems.com/swui/keywords

In the practice of User-Centered Design and Information Architecture for information-focused sites, we reach a point in any project where we need to identify the key words and phrases for the subject domain and the content. We use these terms as the basis for labeling, navigation, categorization, search engine tuning, faceted browsing and filtering. The quality and relevance of the terminology to the user is a critical factor in overall site usability.

Traditionally, we have developed numerous manual techniques to identify, refine and unearth important terminology. There are also many group facilitation activities that we employ to help us refine and gain consensus among subject experts and users.

The good news is that increasingly automated tools are available to help us start to identify important and useful terminology. *"Start to?"* you ask? Yes, no magic wands in terminology identification! Automated tools, such as tag generators, semantic parsers, and concept extraction software do not remove the need for the manual and group activities, but they can make it quicker to get started by identifying important terms which you can then go on to discuss with subject experts.

So what are these tools? Below are five broad categories of things you can do today. I will briefly survey these areas and provide examples, primarily for sites and applications available free online or using open source software. *This is not an exhaustive list* – there are many tools and sites you can use, and new ones appearing all the time.

- Generating keywords and tags
- Extracting concepts and identifying language patterns
- Finding alternative words and phrases
- Borrowing existing vocabularies
- Browsing subject domains

Generating keywords and tags

You've seen the tag clouds and detailed tag lists on many blogs and Web 2.0 sites. You may even have said *"I wish I could implement that for the site I'm working on."* I'm not necessarily promoting the use of tag clouds and lists – they have to fit your users' situations, and there are some challenges with using them effectively. However, they can be a useful way to identify the core terms in your content site, and also give you a sense of how prominent certain terms are in the content… are the large-font-size terms in the cloud the ones you expected to be emphasized?

Examples include:

- **Cloud generators (embedded):** There are a couple of these based around a web service. You paste a small amount of javascript into a web page. The code calls the service, which then displays the cloud in your page. You could add this to a development version (rather than a public/production version) of a site you are designing, to analyze terms in the resulting cloud.
  http://www.tagcloud.com/
  http://zoomclouds.egrupos.net/ (beta)
- **Cloud generators (online):** These sites allow you to enter a URL or copy a block of text into a field, then create a tag cloud on the fly. The results give you an indication of the use of individual terms (not phrases) in your content. The Keyword Density Checker goes a little further, also providing a list of the terms ranked by the number of times they occur, and the density of the terms as a percentage of the overall words on your site.
  http://www.tagcrowd.com/ (alpha)
  http://www.tagthe.net/ (beta)
  http://www.webconfs.com/keyword-density-checker.php

Extracting concepts and identifying language patterns

Particularly with large, older legacy bodies of content, it can be difficult to identify important terminology and understand what information is hidden inside the content itself. Some organizations find themselves with thousands and even millions of documents that need to be cataloged, organized, and then have keywords applied so that they can be found and searched more easily by their users. "Concept extraction" is the field that focuses on this need, and while it is not a magic wand for the problem of categorization, it can be a very useful aid for the IA or categorization specialist. The related field of "entity extraction" is a very close cousin. It is useful for certain domains where the information is changing constantly and identifying relationships between known entities is important (such as people, companies, locations).

There has been a lot of movement in the commercial concept extraction world (companies such as Inxight, Teragram, Endeca, Autonomy, Clear Forest). However, not as many options are available on the web or via the open source community. This isn't surprising, since this is pretty complex technology. It also requires sophistication from the person using these tools – to make sense of the output and get usable value, it often requires "tuning" both the engine parameters and the rules that govern the way output is produced. However, in larger projects the benefits can be worth the investment.

Examples include:

- **Term Extractor:** An online facility developed and hosted by the Linguistic Computing Laboratory of the Sapienza University of Rome, Italy. Once you sign up for a free account, you can upload one or many documents and receive a semantically-derived list of core concepts and terms. Using more advance techniques, you can also "tune" how it parses and extracts terms.
  http://lcl2.uniroma1.it/termextractor (beta)
- **Yahoo Term Extraction API:** Part of the extensive Yahoo developer library now available, you can pass the API a batch of text and it returns significant keywords based on interpreters similar to those used in the Yahoo search engine.
  http://developer.yahoo.com/search/content/V1/termExtraction.html
- **Text Pattern Analyzer:** Developed by the University of Maryland's HCI Lab, this tool is not (yet?) available as an open source toolkit. This is not an extractor. It is an interesting approach to identifying language patterns and comparing language occurrence across multiple documents, and has a great interface. Very useful for seeing the "key" messages arising in a body of content, and evaluating how you might cross-link content at a deeper level.
  http://www.cs.umd.edu/hcil/textvis/featurelens/ (research)

Finding alternative words and phrases

We've used these sites when we get stuck in our writing, but it's worth considering how you can use them to challenge your assumptions about the language you're using on a site. Are you trapped with internal jargon when there are more familiar words people might use instead? Are there different ways of representing your ideas? And, just as important, are there possible confusions or double meanings in the words you are using, which could confuse your users?

Examples include:

- **Thesaurus.com:** A free online thesaurus. This site is part of a larger collection of useful tools that includes dictionary.com and encyclopedia.com.
  http://www.thesaurus.com/
- **WordNet:** Developed and managed by Princeton University, this is a huge vocabulary collection. It is used extensively in research projects relating to natural language understanding/parsing and semantic processing. Increasingly, there are other universities around the world that are producing multi-lingual wordnets that are aligned with the English version.
  http://wordnet.princeton.edu/, http://www.illc.uva.nl/EuroWordNet/, http://www.globalwordnet.org/
- **Merriam-Webster Online Thesaurus:** A big name in the dictionary and thesaurus world has a useful online version, as well.
  http://www.m-w.com

Borrowing existing vocabularies

There's a lot already out there on the Web. Many subject domains have been mapped already, and more are appearing all the time. If there isn't a coherent, useful vocabulary that you are able to adopt from within the project or content set that you are managing, then you might be able to borrow one. You can also use these as a "reality check" on both the terminology and the structure of what you find within your existing content, to see if another published data set aligns with the terminology you've identified.

Examples include:

- **Open Directory project:** A high-level categorization scheme which can give you an idea how your subject domain fits into a larger classification.
  http://www.dmoz.org/
- **GeoNames:** Over 8 million geographical names available under a Creative Commons license, used either via download or direct API.
  http://www.geonames.org/
- **SWOOGLE:** A search engine for millions of publicly-available ontologies and vocabulary data sets for the Semantic Web.
  http://swoogle.umbc.edu/
- **Linking Open Data Project:** A W3C-supported initiative, this group has begun compiling a reference set of large, publicly available RDF data sets for Semantic Web research. There are likely to be many additional uses for such a resource!
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- **MeSH (Medical Subject Headings):** This online classification system is developed by the National Institutes of Health, and is part of the larger Unified Medical Language System (UMLS). It is used for classifying many of NIH's medical publications. The UMLS as a whole contains over 139 separate medical vocabularies.
  http://www.nlm.nih.gov/mesh/meshhome.html

Browsing subject domains

One of the simplest things you can do is look at other places where your subjects are available, and see how the subject is organized and described. How does another public presentation of the subject represent its key concepts? Is it similar to the way you are organizing your content and concepts? There is a very good chance that your users will have some familiarity with other large, popular sites with similar terminology, so it is useful to check and also to get some ideas.

Examples include:

- **Amazon:** Search or look up the top book references for the subject domain you are working on. For each book, look at the subject classification section of the description (near the bottom of the page) to see the classification scheme. Look at the other information that Amazon provides: statistically improbably phrases (identifying distinct terminology in the domain), capitalized phrases (often denoting importance), and even user tags.
  http://www.amazon.com/
- **Del.icio.us:** Navigate the del.icio.us tag lists to see how different terms are related, based on how they frequently occur relating to the same sets of links.
  http://del.icio.us/
- **Cloudalicious:** This is an interesting adjunct to del.icio.us… it provides a graph of a site's tag usage on del.icio.us. If you are working on a public site that is likely to have a presence in del.icio.us already, then you can see a map of how tag usage has changed over time. This may help you tune the messages and categorization on your site. It can be particularly useful if you introduce new terminology on your site (or start posting content relating to a new subject domain), and want to see if the new terms are translating into new tags on del.icio.us, meaning users are beginning to adopt the terms you have introduced.
  http://cloudalicio.us
- **Wikipedia:** There are a number of ways to identify useful terminology in Wikipedia. Browse the extensive list of topics that provide a broad table of contents for the articles. Look up specific subjects of interest and identify highlighted and linked terms. Use the Wiktionary.
  http://www.wikipedia.org, http://www.wiktionary.org