Metadata and the web:

Can automated metadata generation tools help organize the Internet?

Jen Ferguson

IST 616

November 30, 2006

Introduction

Cory Doctorow (2001) famously described the idea of an orderly Internet well-tagged with accurate, usable metadata as a "pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities." Yet even during the earliest beginnings of the web, the drive to organize has been a factor. Tools to help locate information resources began to appear very shortly after the first web browsers did in the early 1990s. Soon after founding the web itself, Berners-Lee founded the WWW Virtual Library, and the appearance of Yahoo!,Webcrawler and Lycos soon followed (Gill, n.d.).

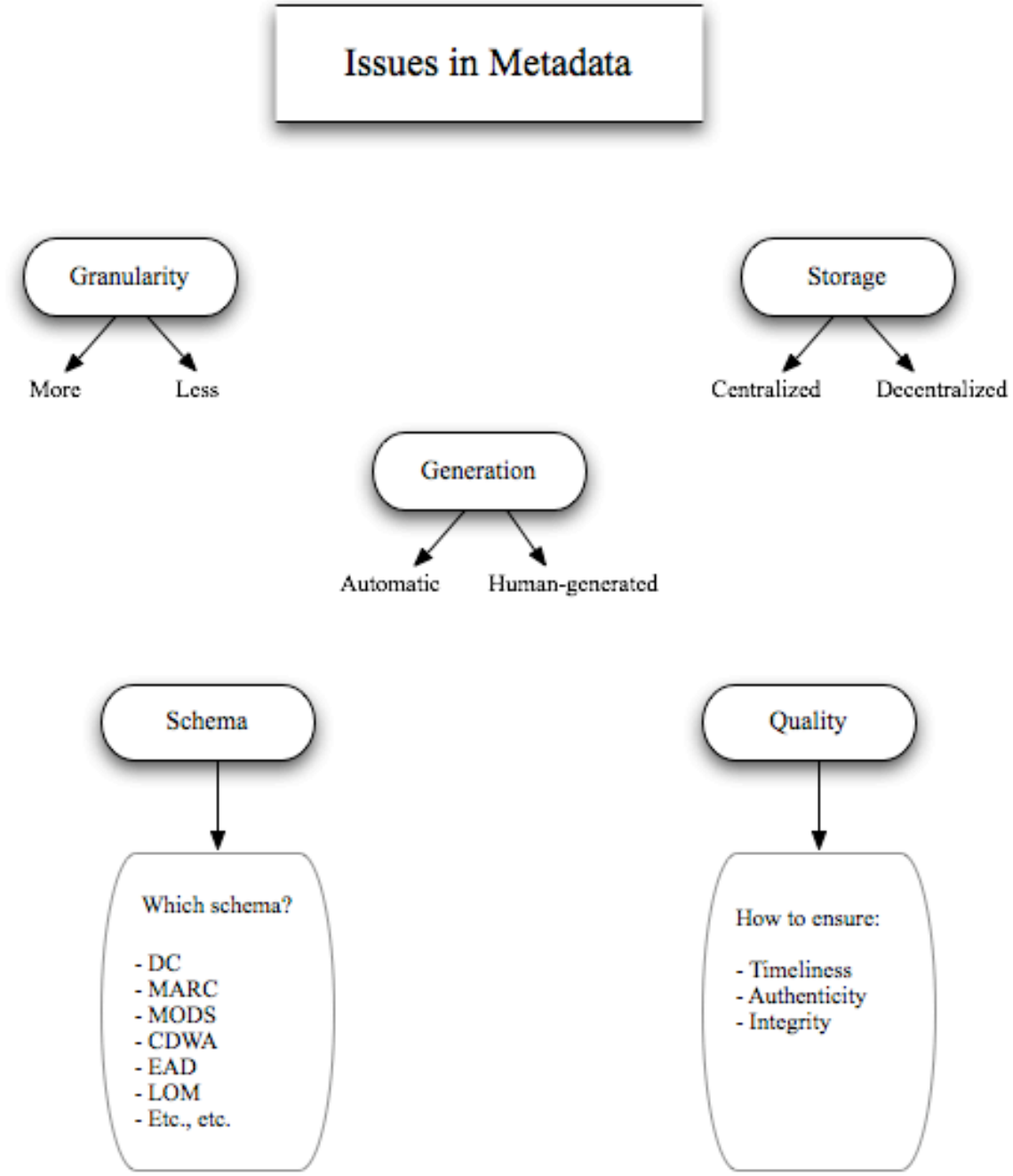
As the web continues to grow exponentially, metadata becomes increasingly important for resource discovery and organization. While search engine crawlers do generate a kind of very cost-effective metadata in the course of trawling the web, more and more of the information they seek is beyond their reach (Gill, n.d.; Duval, Hodgins, Sutton, and Weibel, 2002).

There are other issues to consider as well. Thomas and Griffin (1999) posed an important question: Who is going to do all the work of making metadata for the Internet? They convincingly argue that while the authors of an Internet resource are in the best position to generate metadata for that resource, there is very little incentive for them to do so. Tony Gill (n.d.) agrees:

> Creating good metadata requires time and money, but there is little incentive for content creators to expend much of either on the creation of metadata descriptions, because many search engines don't use them. The metadata that does exist, most of which is created in good faith, is not being used by search engines because they cannot rely on it to provide accurate and faithful descriptions. The missing ingredient is trust, without which the Web's resource discovery cake has a bitter taste.

While there is no one solution to all of these problems, there is general agreement that automated metadata generation tools can help make a dent in the vast amount of as-yet-uncharacterized resources on the Internet.

Metadata generation is just one of the issues involved in metadata for the web. The following diagram illustrates the generation question presented as one part of a broader context.



Inspired by Hunter, J.L. (2003). A survey of metadata research for organizing the web. *Library Trends 52*(2), 318-344.

How automatic metadata generators work

In using a metadata generation tool, the first step is to supply the tool with a method of locating the resource. This is as simple as providing a URL or another web address. The generation tool then uses an algorithm to sift through the content of the resource (source code included) to generate automatically assigned metadata. For example, once the DC-dot generator (www.ukoln.ac.uk/metadata/dcdot) is supplied with a URL, it generates a Dublin Core record of that resource. The resultant metadata can be included in an XHTML/XML document header or housed in a database (Greenberg, 2003).

There are some differences in the technical approaches taken by automatic tools in generating metadata. One technique uses neural networks, employing parallel processing and evaluation of the connections between different items to detect patterns. Certain knowledge bases and rule sets are also used to assist in classifying documents. Chester (2004) offers an example in which documents with a recognizable style of title can be assumed to be members of the 'press release' category. Reamy (2002) suggests that the foremost strength of auto-categorization tools is to very rapidly perform full-text searches of documents and evaluate the resultant word patterns and frequencies. The outcome is then compared with existing categories in a technique called Bayesian analysis. Another approach, called clustering, uses the software to sift through a collection of information resources and identify common clusters of words to find documents that appear similar. Auto-summary approaches are also used; this method scans through documents to pull out sentences that are deemed important. As one example, the first sentences of paragraphs are generally deemed important and are heavily weighted. Yet another approach uses noun-phrase extraction, which generates lists of nouns, or 'things,' included in document collections (Reamy, 2002).

Human vs. machine-generated metadata

Human metadata generators do have a major advantage over automatic metadata generation tools. Despite continuing advances and refinements in the tools, humans still have the upper hand in accuracy. However, relying on human-generated metadata is very time-consuming as well as expensive.

Additionally, some question whether humans are too subjective in their assignment of metadata (Hunter, 2003).

Considering the vast amount of information in need of categorization and resource discovery on the web, it seems clear that automatic metadata generators are of vital importance for organizing this digital information. They are great time-savers and productivity enhancers, and are less expensive than their human counterparts. Like humans, however, automatic metadata generators also have their problems. They still have a high error rate, and at present are much better suited to some applications of metadata than others. Reamy (2002) suggests that automatic generators are ideally suited to situations in which a constant, high volume of documents need metadata assigned to them; that the tools do a much better job of assigning metadata and categorizing documents when those documents are professionally written; and that the categorization required is either of a very general nature or uses a highly technical vocabulary.

In his 2002 paper, Reamy cites a statistic that automatic metadata generation tools can achieve about a 90% accuracy rate. This sounds rather impressive, until the realization dawns that one out of ten documents returned as search results will be incorrect. When used in combination with evaluation by librarians or other information professionals, however, that accuracy rate jumps to 99%. Little wonder, then, that there is much agreement in the field that a hybrid system combining machine-based tools with human expertise is the best approach to accurate metadata generation.

Summary/Conclusion

From my readings, several clear consensus points emerge:

- Metadata is expensive to generate (in both time and money) but very valuable.
- Automatic generation tools are useful and necessary.
- These tools are best put to use for more routine aspects of metadata generation.
- Humans are still an important part of the metadata generation process, particularly in the more intellectually challenging metadata applications such as subject descriptors.

There were some variations on this theme. For example, Thomas and Griffin (1999) suggest that commercial indexing services have the most incentive for deploying good metadata. Mitchell (2006) is an

advocate for library-developed metadata generation and management systems as the best choice. Stoica and Hearst (2004) promote the benefits of exploiting existing lexical hierarchies for the most accurate 'nearly-automated' metadata generation.

I was surprised by the degree of unanimity I found in my research of this topic. Everything I read enthusiastically supported a hybrid approach to metadata generation, utilizing both human and automatic generators. There was agreement that automatic metadata generation is both important and necessary, but also that it cannot be the only solution; humans are still needed in the equation. While my personal experience with metadata generation tools is not yet extensive, I came to the same conclusion after using the DC-dot metadata generator for an Internet resource during the skill workshop for this course. The generator did a nice job of extracting many of the more basic details from my web site (such as the title, creator etc.) but the best results came after I looked over the results and edited some of the more complicated fields.

I believe the future holds much promise for automatic metadata generation tools. As natural language searching capabilities continue to develop, further refinement is bound to improve the accuracy of these tools. The evolution of metadata generation for multimedia resources will be an interesting process to watch. As the world grows ever-smaller with the increasing reach of the Internet, I am in agreement with the assessment by Duval et al. (2002) that multilingualism and multiculturalism will be grow in importance, impacting many aspects of metadata from date order to creator name order to even the direction the text is written. Finally, increasing involvement by librarians and libraries in developing these tools themselves as discussed in Mitchell's 2006 paper is a very promising direction. The combination of 'modern' techniques such as full-text search capabilities with tried and true library techniques like applying controlled subject headings bodes well for the future of automatic metadata generation.

Annotated Bibliography

Duval, E., Hodgins, W., Sutton, S, & Weibel, S.L. (2002). Metadata principles and practicalities. *D-Lib Magazine 8*(4). Retrieved November 20, 2006, from http://dlib.org/dlib/april02/weibel/04weibel.html

In this comprehensive general overview, the authors discuss the confusion about how best to include metadata in information systems. They split their ideas into two broad categories: principles, or ideas held in common by all metadata standards, and practicalities, problems that arise when moving from theories to real-world applications.

Their discussion of principles touches on the concepts of modularity (in both namespaces and metadata), extensibility, refinement, and multilingualism. Practicalities mentioned include application profiles, syntax and semantics, association models, identifying and naming metadata elements, metadata registries, thoroughness of description, mandatory vs. optional elements, and subjective and objective metadata. The article closes with discussion of automatic generation of metadata.

Gill, T. (n.d.). *Metadata and the world wide web*. Retrieved November 19, 2006, from http://getty.edu/research/conducting_research/standards/intrometadata/metadata.html

This is an excellent general treatment of metadata on the web. It includes some historical perspective, as well as discussion of the difficulties of finding information on the Internet, the problems inherent in methods used by search engines, and explores the question of whether the web can and should be catalogued. The bulk of the article concerns metadata: its applications and its issues, tools and standards.

While the article was definitely helpful to me in writing this paper, its usefulness stretches far beyond this particular topic. I found this piece well-written and interesting, and will definitely be keeping it in my library in the expectation that I'll be referencing it again in the future.

Greenberg, J. (2003). Metadata generation: processes, people and tools. *Bulletin of the American Society for Information Science and Technology, 29*(2), 16-19.

Jane Greenberg presents an outline for the successful generation of metadata. As the title suggests, she advocates for the importance of integrating people, processes and tools in this task. She compares the classes of metadata generation tools (templates, generators and editors) as well as the different categories of human metadata generators, which run the gamut from professional metadata generators to subject knowledge enthusiasts.

Greenberg closes with a survey of some metadata generation research projects, including one called Breaking the Metadata Generation Bottleneck, which I was surprised to discover was headed by SU's own Liz Liddy.

Hunter, J.L. (2003). A survey of metatdata research for organizing the web. *Library Trends 52*(2), 318-344.

Hunter details some of the advantages and disadvantages of using metadata. This article also covers research areas the author feels will be most important over the next several years. These include XML (including namespaces), Semantic Web/interoperability issues, metadata harvesting and the Open Archives Initiative, metadata for multimedia, rights metadata, automatic metadata generation and auto-categorization, search engine research and development, multimedia/graphical presentation of results, and metadata for customization. The piece concludes with current key issues in the field.

Mitchell, S. (2006). Machine-assisted metadata generation and new resource discovery: Software and services. *First Monday 11*(8). Retrieved November 27, 2006, from http://www.firstmonday.org/issues/issue11_8/mitchell/index.html

I wish I'd had more space to delve into this paper. Mitchell writes about iVia and Data Fountains, initiatives he directs that provide open source software to the library world. The Data Fountains project includes a metadata generation utility both for natural language searching of key phrases and a utility for controlled subject heading generation such as LCSH. Other features include metadata extractors, full-text

ID and extraction, and an Internet resource discovery service. These are available both in semi-automated and fully automated modes.

These projects are somewhat unique in that they are not only intended for use in the library setting, they are also created by librarians. The melding of full-text search capabilities with conventional library classification schemes, designed by and for librarians, seems quite a promising direction for the field.

Reamy, T. (2002). Auto-categorization: Coming to a library or intranet near you! *Econtent 25*(11), 16-18, 20-22.

This piece provides an overview of how automatic metadata categorizers work. Reamy describes the use of Bayesian scanning of the word patterns found in full-text documents to assign documents to categories. He also discusses clustering document collections, auto-summary functions that search for important sentences in works, metadata generation via first categorizing a document, then searching for keywords to match the assigned categories, and noun-phrase extraction,.

The article includes the interesting historical tidbit that these auto-categorization tools got their start in the news and content provider arenas, and goes on to highlight the situations these auto-categorization tools are currently best equipped to handle.

Thomas, C.F., & Griffin, L.S. (1999). Who will create the metadata for the Internet? *First Monday 3*(12). Retrieved November 16, 2006, from http://www.firstmonday.org/issues/issue3_12/thomas/index.html

This arresting article first drew my attention and attracted my interest in writing on this topic. Thomas and Griffin call attention to the fact that while many are debating the merits of metadata and dissecting the finer points of schemas, the question of who (or what) will be doing the work of generating metadata for the Internet is not being widely addressed.

The authors make the compelling argument that while users are often the best generators of metadata, there is very little incentive for them to do that generation. They make a case for commercial indexing services being in the best position to profit from the assignment of metadata, and also see a role for use of metadata generation tools in “taming the Internet”.

Other references

DC-dot Dublin Core Metadata Generator. (n.d.). Retrieved November 25, 2006, from http://www.ukoln.ac.uk/metadata/dcdot/

Chester, B. (2004). Auto-categorization and records management. *AIIM E-Doc Magazine, 18*(2), 16-18.

Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia [Electronic Version]. Retrieved November 25, 2006 from http://www.well.com/~doctorow/metacrap.htm

Stoica, E. M., & Hearst, M.A. (2004). *Nearly-automated metadata hierarchy creation.* Paper presented at the Companion Proceedings of HLT-NAACL'04, Boston, MA.