

## **Vertical Search, Meta Information, and XML**

### Reaction Paper 1

Christian Montoya, CS 431, Feb. 22, 2007

Published at <http://www.christianmontoya.com>

Modern horizontal search engines that cover a wide range of web documents and attempt to index all of the web pages and media available on the Internet are lousy. As much as these search engines attempt to improve their algorithms and build better representations of the text that they mine, they cannot provide any meaningful or useful catalogs of their indexes or present the end user with effective ways of refining queries. This paper proposes new ways to use XML with publishing platforms and information networks to complement text and media search engines with useful meta data. The hope is that by providing meta information about web content with XML, the modern information problem, that of users being unable to find the information they need on the World Wide Web, would be one step closer to being solved.

In *The Intellectual Foundation of Information Organization*, Svenonius (2000) describes four basic bibliographic objectives: to find entities that correspond to the user's stated search criteria, to identify an entity, to select an entity that is appropriate to the user's needs, and to acquire or obtain access to the entity described. Modern day search engines, by virtue of their ranking algorithms, mining capabilities and the nature of the web, are able to fulfill these objectives to varying degrees. They are able to offer results that are as good as the user's query, identify results based on very limited meta information, select appropriate results based on rankings, and direct users to the right URLs that allow them to access those results. The ability of search engines to meet these objectives, being entirely machine-based and unable to develop any cognizant idea of the content they are mining, is entirely based on the

amount of meta information available for web content. Svenonius also describes a fifth objective, the navigation objective, which is impossible for search engines to fulfill without a wealth of clearly defined meta information. The navigation objective defines the means by which information in a collection is organized in such a way that users who do not initially know exactly what they are looking for or how to present their queries can find the results they need simply by navigating through the collection and following the organization to discover the results. Meeting this objective in a search engine requires significant information about the content in the search engine index and a serious attempt to catalog that content, however simple the cataloging criteria may be. It is without the fulfillment of the navigational objective that the modern information problem remains unsolved, with search engines providing a method for users to access information on the web that is insufficient.

Furthermore, horizontal search engines, that is, search engines which attempt to index as much content as possible, cannot possibly index enough content on the web to provide users with a decent snapshot of all the content that is actually available. In the year 2000, a study at BrightPlanet (Bergman, 2001) quantified the size and relevancy of the "deep Web" as 400 to 550 times larger than the commonly defined World Wide Web, with 7,500 terabytes of information and nearly 550 billion individual documents. At the time, none of this content was visible to current search engines. Seven years later, with the proliferation of broadband Internet usage and the availability of visual, auditory, and interactive media as well as quick publishing systems that automate syndicated content, it is easy to understand why horizontal search engines can only skim the "tip of the iceberg" when answering queries. Even though none of the major search providers are willing to publish the size of their indexes, it is clear that no search engine would ever be capable of offering access to any significant fraction of the content available on the web. Even if they could, they would only truly be useful if they could ensure that the first results returned for any query were the absolute best results

possible. Therefore, when faced with the problem of making the bulk of information available on the web to users, horizontal search is not a solution.

In contrast to horizontal search engines, vertical search engines narrow the index they provide according to categories or relations in order to provide users with refined results and less noise. A degree of vertical access exists on modern search engines that offer a variety of secondary engines such as image, video and web log search, but unfortunately modern text-mining is incapable of gleaning sufficient meta information from web documents to actually identify them, much less non-text data such as images, audio and video. From its inception, HTML offered ways to associate meta information with web documents, but these methods were limited and far too easy to misuse. It was because of widespread misuse that search engines began ignoring meta information entirely, and that is what has brought us to this modern age of algorithmic text-mining information retrieval. In order for search to improve, and for vertical engines to fulfill their purpose, meta information, and more importantly, cataloging, needs to be brought back into the search equation.

In *Metacrap*, Cory Doctorow (2001) argued that meta information fails for a number of reasons. According to Doctorow, people are bound to lie for financial gain, are too lazy or stupid to classify things properly, or simply cannot agree on the right cataloging methods. Search engines have done good work to abandon the blind faith of meta search in the past and develop algorithms that work independent from human involvement. The current search model, however, exists much like the original query in, results out model described by Ramana Rao (Rao, 2004). Even with implicit meta information such as "page rank," the means by which modern engines such as Google rank results based on their popularity, the combination of poor queries and few ways to understand documents results in a broken model that does not work often enough.

It is because of this failure to satisfy that users have turned to new options. They rely

on social bookmarking and the wisdom of other users, whether through web logs, forums or other communication platforms, to acquire the results that they need. These resources, provided by humans, are often better than what search engines can provide because humans possess the innate ability to identify meta information for content even when it is not explicitly provided. Moreover, new ways of adding meta information to web content have arisen, one example being microformats (<http://microformats.org>), which build upon existing XML and XHTML standards to provide humans and machines with meta information relating to web content such as calendars and reviews, among many other things. Microformats have risen in popularity because the need for them is huge for both developers and end users, and XML is what made them possible. They are easy to use because they can often be generated automatically by the same software used to publish the web content in the first place. They also address the issue that basic meta capabilities provided in HTML are not sufficient to accurately identify the wealth and variety of data on the web.

What search needs is a return to meta information as a means to complement the existing text mining, relevance and popularity algorithms, but to make this possible requires a new approach that is not dependent on publishers creating the information by hand. Expecting users to properly catalog and identify the content that they publish on the web will only result in a return to the problems described in *Metacrap*. Today, however, it is possible to rely on the same software used to power and manage websites and information systems to generate useful meta information as well. Content management systems, for example, possess the ability to automate the entire publishing and cataloging process, and these systems power the majority of sites on the web today. These systems already publish meta data automatically for syndicated content, such as web log entries and news articles, as well as for entire websites. One example of this would be Wordpress, which powers my design and technology web log. Wordpress publishes both an RSS 2.0 XML feed of the latest

10 entries with full RSS-compliant data about each one, as well as an XML site map that identifies the various sections of my website for search engines as well as estimates of how often each section should be revisited. These tools exist and function outside of my control; I cannot change the way they operate and they require no effort from me in order to generate meta information about my website. It is because of automatic XML generation in web log software that web log search engines have been so much more effective than traditional text mining engines in returning relevant and recent content for natural queries, but even those are still far from their potential.

Another large source of web content and information, social networking sites, also possess the ability to provide expert cataloging information about the content in their networks. A video on YouTube, for example, contains some degree of implicit file data and user provided information that could be republished as meta information. There still exists the issue of users lying about the correct information for their videos, especially when it comes to open systems such as "tagging," but a social network like YouTube also possesses a large base of users that can flag and collectively revise or ban content that is categorized incorrectly. Therefore, the community can collectively regulate the content they publish. The YouTube platform could then collect this implicit file data and user provided information and generate XML meta data for search engines to crawl, which would allow video search engines to provide more narrow and relevant results with some notion of what each video actually contains. By automating the process and relying on community self-governance over individual control, traditional problems with user-provided meta information are overcome.

With proper meta information available for web content in XML, vertical search engines can refine their indexes by looking for specific cataloging information about the content they crawl. The exact cataloging methods used are not essential to this proposal; it would simply be necessary that the methods have widespread adoption and considerable

flexibility. An example of an XML cataloging method would be applying location information to photographs contained on photo sharing sites like Flickr (<http://flickr.com>) and Zoomr (<http://beta.zoomr.com/home>) and allowing crawlers to retrieve this information and use it to provide vertical search categories based on cities, regions, and countries. With the help of automated XML cataloging, a search for "springfield pictures" with the intention of finding photos of Springfield, Florida, which would probably return Simpsons screen captures intermingled with photos from Springfield in all 50 states in a current photo search engine, could be applied to a State and City filter in a meta-data-dependent search engine that would restrict the results and return far less noise. By making this meta information available in lightweight XML files rather than embedded in complete web pages, the cataloging information would be easy for the software to maintain and quick for the search engine to crawl.

With the help of XML cataloging, web content would begin to resemble the information workspace model described by Ramana Rao. More advanced search and filtering capabilities, and even ways to browse indexes, would allow users to better find the content they are looking for and decrease the reliance on plain-text queries. As for regulating the cataloging of web content across billions of web sites, the problem already exists in the form of "spam" websites and methods of gaming the "page rank" algorithm. Major search engines currently offer ways for users to report misleading content, and they could refine their algorithms to compare XML catalogs to the actual site content. Therefore, it is possible for search engines to rely on meta information to effectively index web content and provide vertical search capabilities to users. With such capabilities, the navigational objective would actually take form on the web, making the right results easier for users to find.

## References:

1. Svenonius, Elaine. *The Intellectual Foundation of Information Organization*. Cambridge, Mass.: MIT Press, 2000.
2. Bergman, Michael K. *The Deep Web: Surfacing Hidden Value*. *Journal of Electronic Publishing*, 7(1), August 2001.
3. R. Rao. *From IR to Search, and Beyond*. *ACM Queue*, 2(3), May 2004.
4. Doctorow, Cory. *Metacrap*. Available on the web at <http://www.well.com/~doctorow/metacrap.htm>. Version 1.3:26, August 2001.