SEARCH ENGINES AND METADATA—CURRENT SITUATION

Metadata standards and their application in the digital environment will become crucial in the efficient harvesting and retrieval of information by search engines. By applying metadata to digital objects and collections at the point of creation, librarians provide the mechanism for the future organization and retrieval of this information both internally through their own information retrieval tools and externally through automated harvesting and interoperable searching initiatives.

Since 1995 metadata standards have been aimed at helping Web users achieve better precision and recall of information when using Web search engines. The use of metadata applied to Web resources during the creation of the resource, or applied by a third party such as information organizations or services, provides better authenticity of information and organization than the current method used by most search engines to index Web information—that is, to send out automated spiders and bots to harvest information from the HTML code and index this information.

The problem with this method of harvesting was that many Web resource creators would stack or stuff important keywords on the front page of the resource or in the HTML code behind the resource. This approach provides that resource with a higher ranking in the search engine, making the resource appear more frequently and on the first page of search results when users initiate a search.

For additional statistics see **Search Engine Index,** http://searchenginewatch.com/reports/seindex.html

Supporters of metadata standards point to this inefficiency, and in particular the high number of irrelevant retrievals and low precision/recall ratio of Internet searches, as why search engines should adopt and search for metadata as their primary indexing criteria.

Why spend so much time trying to make search engines better and of more use to Web searchers? More than 75% of Web users use search engines to traverse the Web, according to the Search Engine Index.

Commercial Web sites, as well as the search engines themselves, should make their information indexing and retrieval methods more efficient. Metadata initiatives can help. Dublin Core was especially designed to be a low-maintenance, common-denominator standard for Web resource discovery.

Have search engines seen the light regarding the incorporation and indexing of metadata by their harvesting spiders and bots? The quick answer is no.

Search engine companies, as well as many metadata creators, realize that the issues of authentication and reliability of metadata are still unresolved and are only a continuation of the current stacking and stuffing problems found within HTML coding. Finally, many search engines companies have spent much time and money making their indexing systems more efficient using the current methods of harvesting, so they are unconvinced that searching metadata is in their best interest.

Does the application of metadata into digital objects and collections really create better search and retrieval on the Web? If search engines do not use metadata as a primary criteria for relevance and precision and do not intend to do so, then why use metadata at all? The expertise of those in information organizations becomes crucial. Librarians already know why people use libraries, and why they access and browse collections—because librarians add value to that information, which in turn assists people in retrieving the information they need.

Although the Web environment enables a want-it-now-and-fast attitude, users have found that what they receive in return is a pile-it-up-and-dumb-it-down list of information junk in return. The popularity of Yahoo is that it looks like a search engine up front, but in fact it is a human-made index constructed similarly to indexes found in traditional information organizations. But this type of human-constructed index still does not address the concerns and problems in the long term regarding the efficiency and accuracy of information retrieval of search engines.

The deeper problem is that search engines only search a certain portion of the information universe the Web enables. Research indicates the Northern Light search engine is the most efficient at searching the Web information universe, but their total percentage of coverage is only 16%, and they serve the corporate market exclusively. Research also indicates that, putting all search engines together and eliminating retrieval of identical Web sites, only 21% of the Web has ever been indexed at any one time.

The deep Web is not being indexed by search engines, and the deep Web is where most of the relevant and value-added knowledge is available.

In addition, certain package and HTML graphics formats hinder indexing by search engines. The best known of these are PDFs and HTML frames coding. Google is now indexing PDFs, but the only way to index PDFs is to include a wraparound metadata package for the search engine to access. Although frames graphics look good and catch the eye, every page within that resource has the same URL, so the deeper the information is embedded, the more that information becomes unindexable.

A **wraparound metadata package** wraps around the PDF, allowing the PDF to be searched by search engine bots and spiders.

Splash pages (pages that precede access to the actual content of a Web site and are only graphical displays) are another example of graphics that catch the user's eye, but provide little to no information for search engine harvesting mechanisms to retrieve. In fact, many search engines do not index a Web resource with a splash page.

If metadata is applied to this information at the front end of resource creation (that is, in the digitization of objects and collections), then embedded metadata become the most efficient means for retrieving information in automatic harvesting endeavors. Instead of trying to convince search engines to adopt metadata standards as their primary strategy for indexing, librarians need to embed metadata in digital objects and collections to efficiently and effectively provide the means by which these resources can be indexed and retrieved in future harvesting protocols.

This work is already happening with the Open Archives Initiative (OAI), where the correct application and standardization of metadata elements in digital objects have proved to be the most efficient and effective means of automatically accessing, indexing, and retrieving digital information. OAI is quickly moving to provide this harvesting mechanism as a free and open-ended service to the world. It is a much-needed interoperability element lacking in today's search engine environment,