Le Yang, June 16, 2016

Ensuring Balance when Evaluating Search Engines

In this issue of JWL, Le Yang describes a study that examined how effective search engines were indexing an institutional repository (IR) in DSpace. The purpose of this letter is to clarify some of the finer points of search engine optimization for repositories and give some context for the study that may need to be presented again. Also the letter is to address the concerns raised in the conversation, facilitated by the editor Hannah Rempel, between Le Yang, Kenning Arlitsch and Patrick O'Brien.

Yang's article was not claiming that the results would be true of all IRs. It is, instead, an independent exploratory study on a local IR of how Dublin Core and DSpace (one of the most common combinations used by many institutions) faired with search engine crawlers with no additional modifications to the system or metadata than what the institution was using to begin with. This kind of independent study is important to test the waters with search engines regarding what they will and won't work with.

In his literature review, Yang cites an article published in 2012 by Arlitsch and O'Brien, titled "Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar." That article made specific recommendations on how to make indexing better specifically with Google Scholar. Readers who are interested in procedures to make their IRs work better with Google Scholar should read the Arlitsch and O'Brien article on recommendations on metadata schemas, file sizes, HTML tag, etc., and follow their suggestions to make digital content indexed by Google Scholar.

Yang's study, far from being contradictory, agrees with Arlitsch and O'Brien in that an unmodified IR may be difficult for Google Scholar to index, but adds the information that the main Google index can find the content just fine with no modifications which is a notion needing further research in the future.

More importantly, Yang's exploratory study investigates how four independent search engines, including Yahoo, Bing, Google, and Google Scholar, interact with the local DSpace-based IR and Dublin Core metadata schema. As a researcher who wanted to carry out an objective study, Yang chose not to follow Google Scholar's guidelines as recommended by Arlitsch and O'Brien to implement Google Scholar friendly strategies to the digital collections. Doing this will give indexing favoritism to one particular search engine, Google Scholar in this case, and may affect the search results from the others.

It has been mentioned that the sample size in this study is small. This was done on purpose to make the exploratory study manageable and to reduce the effect of the study on users of the IR. Moreover, it is unlikely that more items in the study from the same IR would yield different results. Yang addressed this issue in the article, that sample size does not alter the search engine's preference on indexing metadata or PDFs. A larger sample size will not change the results on items which have been indexed and have shown up in the research. Both metadata and PDFs were indexed suggesting that sample size was not an impacting issue.

It is also recommended that researchers do a test on their IRs like the one conducted in Yang's article to get a baseline on how search engines are interacting with the content, so when researchers make changes, they can gauge the effect of those changes. Researchers should also publish the results so that the community can start gauging the impact in aggregate. Researchers should note, though, that when

conducting research they should not follow one particular search engine's indexing guideline, because doing so may yield a biased result.