



Discovering Open Access Content

Tamir Borensztajn, MLS
VP, Discovery Strategy

Joseph Henry, the first Secretary of the Smithsonian, wrote that “the worth and importance of the Institution is not to be estimated by what it accumulates within the walls of its building, but by what it sends forth to the world.” (<http://siarchives.si.edu/history/exhibits/henry/joseph-henrys-life>). With the advent of the Internet, sending information “forth to the world” has become easier than ever, but is still not trouble-free. The question today is one of discoverability. How can we ensure that those in need of relevant information can readily find it?

It goes without saying that the subject of discoverability takes on additional significance in the context of open access. Open Access research is only impactful, after all, if it gets found. This is where a commercial discovery system (a.k.a., discovery service, discovery layer, etc.) should come in. A discovery system creates a central, searchable index of an institution’s holdings, including open access content, which is essentially part of everyone’s holdings. What we, as consumers of information, require in any discovery system is the best way to match up an article with a person’s specific research needs. While many systems lay claim to the best practices in this realm, success is truly a function of the underlying approach to indexing and relevance ranking.

Relevance ranking is a complex science that should involve many developers and intricate algorithms to determine how articles get classified and subsequently weighted in search results. This ranking system should be sophisticated, so the user doesn’t have to be. The discovery system vendor should be open in its approach to remove the mystery for customers and content providers. While free search engines do not share their inner workings with the world, commercial services should do so. Sharing the recipe does not mean that the work is easily duplicated by competitors. A search for *relevance ranking* on Google should immediately answer questions on how each system works. Try this search some time.

By creating high quality indexing for open access journals, next generation relevance ranking systems can ensure that the most relevant articles appear at the top of the user’s result list. But even then, most searches have many relevant results, so how do we ensure that the most valuable articles are presented to the end user? Usability testing provides us with many of these answers. We see that, across the board, when all things are equal in terms of relevance, the academic end user will opt for the 2014 article over the 1974 article, the four-page peer-reviewed paper over the two-sentence news blurb, etc. Users want discovery systems to make these types of value judgments for them.

What’s more, users want to be assured that they have access to valuable open access content irrespective of any commercial interests or affiliations. This means that no discovery vendor should

employ a relevance algorithm to favor content from a publisher who is a commercial partner. Likewise, it would be wrong for a content aggregator who is also a discovery vendor to employ no ranking algorithm whatsoever. No ranking algorithm may inadvertently result in bias in the discovery system as certain content with greater frequency of publication (e.g., newspapers, trade periodicals, etc) will “win”. These sources truly dwarf the number of open access scholarly articles, meaning that this subgroup would be duly underrepresented.

One thing is certain: basic value judgments must be made for content to be fully discoverable. Yet we must be extremely careful with the decisions made in this area. Extensive testing of all value judgments is entirely critical, and should be open for all to understand. Improvements should be a collaborative effort between discovery system vendors and their customers.

Another imperative consideration in discoverability of open access content is the scholarly value of the research and its overall quality. It is necessary to vet open access content for the end user. We are now at a point where some of the best journals in the world are open access; however, some of the worst journals are also open access. We do not want a few bad apples to spoil the wagon. Some open access journals have concocted editorial boards, include author misconduct to the extent that the journal suffers from repeated cases of plagiarism, do minimal or no copyediting, and /or are obvious pseudo-science. These publications corrupt research and are bad for the academy. Each discovery system vendor should keep its own list of these journals. If the vendor doesn't have the resources to do so, it can refer to academia's publicly available watchdog list for such publications (Beall's List). (<http://scholarlyoa.com/publishers/>). While this is a controversial topic, open access advocates must work together to acknowledge that a problem exists and collaborate to avoid allowing these publications to tarnish the fantastic work being done by the majority of open access publishers.

As a research community, we must support the indexing of open access journals, the creation of value ranking algorithms in discovery systems that do not smother open access journals, and the vetting of open access journals in order to expose the large majority that are of good or great quality.