The Case for OAI in the Age of Google

From "The case for OAI in the age of Google," SPARC Open Access Newsletter, May 4, 2004.

http://dash.harvard.edu/handle/1/4455494

Why don't more faculty deposit their eprints in open-access, OAI-compliant (OA-OAI) archives? This is a mystery. Two explanations we can rule out right away are opposition to open access and opposition to OAI [Open Archives Initiative, short for OAI-PMH or Open Archives Initiative Protocol for Metadata Harvesting] metadata sharing. These never come up when faculty are asked about their archiving inertia, which only makes the mystery even more puzzling.

When asked about archiving inertia, some faculty say that putting an eprint on a personal web site is just as good as putting it in an OAI-compliant archive. Google will find an eprint on a personal web site and make it visible to those who might need it for their research. Let's look at this one more closely. *Is* Google just as good? How strong is the case for OAI archiving in the age of Google?

For this purpose, let me use the name "Google" to represent not only Google itself but any Google rival or future iteration of Google that improves on Google's famously effective relevancy algorithm and wide scope. In short, let "Google" be our name for the state of the art in indexing by mainstream search engines.

So, is Google good enough? If not, why not?

(1) If we only care about open access itself, then it's true that putting an eprint on a personal or institutional web site is good enough. It's open access. […]
(2) The OA-OAI proponent might concede that eprints on personal web sites can be OA. "But OA-OAI archiving enhances visibility more than Google indexing does."

The Google reply: This may have been true once, but it's less true or untrue today. There are two reasons why: Google is very good and getting better, and many more

people turn to Google before they turn to OAI search tools. The second reason is peculiar. […] Even if OAI tools would do a better job than Google, if they were as popular as Google, Google's surpassing popularity gives it a self-nourishing advantage. […]

(3) This actually answers another argument that might be made for OAI archiving, but let's make the argument explicit anyway. "Scholars doing serious scholarly research look in specialized scholarly tools and resources before they look in Google."

The Google reply: Again, this might have been true once, and perhaps it ought to be true now, but either it's becoming untrue or it's already untrue.

Working researchers certainly do use Google even if they also use specialized scholarly tools. Moreover, unfortunately, in the period before scholars used Google for serious research, they weren't using OAI tools instead. Google and OAI tools are both rising in usage.

Two years ago (April 2002) a study by DK Associates showed that professional analytic and organic chemists turned first to ChemWeb and second to Google. It's impressive that Google occupied such an exalted position among such serious researchers that early in its evolution. The same study showed that chemists in management and development positions used Google first and ChemWeb second.

http://web.freepint.com/go/newsletter/109#feature

Since then, Google has improved its algorithms, its index size, and its popularity—and Elsevier has decided to discontinue ChemWeb. I haven't seen a more recent study, but I wouldn't be surprised if Google was #1 among working chemists today.

In his February 2004 keynote at the NFAIS annual meeting (p. 8), John Regazzi reported, "In a survey for this lecture, librarians and scientists were asked to name the top scientific and medical search resources that they use or are aware of. The difference is startling. Librarians named Science Direct, ISI Web of Science, and Medline, while scientists named Google, Yahoo, and PubMed (librarians also named PubMed)." Regazzi is Elsevier's Managing Director of Market Development.

http://www.nfais.org/RegazziFinalMilesConrad2004.pdf

[…]

(4) "Archiving will give an eprint a permanent or persistent URL." Compared to eprints on personal web sites, eprints in OAI archives rarely move. When scholars change institutions or retire, they usually change web sites, with the effect of breaking links that point to their work e.g., in search engines, bibliographies, footnotes, and other indices around the world. This is a reason to favor archives over personal web sites.

Google reply: True, but Google has a large and useful cache that greatly mitigates the damage of link rot.

Archiving will give an eprint other kinds of longevity, not just URL longevity. Those who maintain OAI-compliant repositories take steps to assure long-term access and preservation. […]

(5) "OAI-compliant searching tools refresh their indices faster than Google."

Google reply: But this is not quite true. Google refreshes its index for different kinds of content at different rates, and assigns a slow rate to most scholarly pages. But eprints at sites that Google already rates highly are refreshed at a much faster rate. On the other side, the refresh rate at OAI-compliant data services is up to the service providers. At least this means that when we want to refresh the index often, we can do so, and we needn't hire expensive experts in "search engine optimization" in order to scam the Google index in a way that might not work next week.

(6) "OAI tools rest on a standardized metadata schema and therefore support field searching (e.g. on 'author' or 'title')."

Google reply: True, but the Google syntax does a lot of this and over time will do a lot more.

Here's a variation on the OAI argument: if users search for articles by their citations, rather than by content-based keywords, then OAI tools will help them more than Google will. I owe this argument to the EPrints Handbook.

http://software.eprints.org/incoming/lac/overview.php

The Google reply: It's true that OAI tools will provide better visibility to those who search by citations. But talented Google searchers will prefer to search by content-based keywords, not by citations. If they do, then they will likely find the same articles by a different route, though they will be combined with all the other articles that also satisfy the keywords. Insofar as the size of the hit list is a problem, see the next OAI argument.

(7) "OAI archiving reduces information overload." When you search across OAI-compliant archives for research literature, you find only research literature. But when you search in Google, you get commodities with the same names, popular literature on scientific topics, scientific name-dropping, crackpot hallucinations, and much more that you definitely don't want.

Google reply: This is true, but it overlooks the Google relevancy algorithm. In Google, you may get more hits than you could ever scan, and many of them will be worse than useless, but Google's PageRank algorithm does a pretty good job of putting the ones you want near the top. Just as it doesn't matter how deep the ocean is, as long as you can swim, it doesn't matter how many hits your search returns, as long as the ones you want float to the top. Moreover, skillful users know how to tweak their search strings to narrow the results and improve their relevancy. Finally, remember, the Google

algorithm (in fact and ex hypothesi) is improving all the time. We don't have to say that the Google algorithm is perfect, merely that a good algorithm can neutralize much of the advantage of a smaller or more focused index and that this one is good and getting better.

Judge for yourself. Here are some terms from different academic fields and their Google hit tallies as of April 18, 2004. Run some of them and see whether any non-academic sites make it near the top of the list. Then tweak the search to refine the list. "Poincare conjecture" (2.8 thousand), "third-wave feminism" (3.9 thousand), "proto Indo-European" (12 thousand), "categorical imperative" (25.4 thousand), "valence electron" (27.2 thousand), "battle of Hastings" (39.7 thousand), "collateral estoppel" (46.9 thousand), "obsessive compulsive" (304 thousand), "black hole" (2.7 million), "inflation" (4.9 million), and "protein" (26.8 million). The general terms toward the end of the list get the most hits. But it's easy to conjoin them with other terms in order to reduce the hit list and improve relevancy. For example, try "black hole" plus "event horizon" (41.8 thousand), "inflation" plus "junk bond" (6.6 thousand), or "protein" plus "chirality" (47.8 thousand).

On the other side, any improvements that come to the Google algorithm could also in principle come to the OAI search tools. That would give the OAI tools a twofold strategy for reducing information overload—intelligent sorting and smaller or more focused indices. But even then, Google could claim a twofold strategy for finding what you want—the same intelligent sorting but yoked to a larger and more wide-ranging index. In short, the same small OAI indices that some cite as an advantage in reducing overload can always be seen as limitations on the search for what you want.

Here's a variation on the same OAI argument: "If you're searching for an unusual author name or term in Google, you'll probably find what you want. But if you're searching for a common term or name, then OAI searches will probably shorten your search."

I owe this argument to a participant in David Prosser's workshop on filling OAI archives at the CERN OAI meeting in February—a participant whose name unfortunately I do not know. If you are searching for "John Anderson," "piano," or "chess," Google will be less useful than if you are searching for "Spiro Agnew," "sackbut" or "43-man squamish".

(8) I'm out of OAI virtues that might surpass Google virtues. Are there any Google virtues that might surpass OAI virtues? Here's one: a gigantic index. But as we just saw, this advantage competes with the advantage of the smaller and more focused OAI indices. For some searches, a wide scope (plus a good relevancy algorithm) is more useful than a manageable hit list (plus a good inclusion policy), while for other searches the reverse is true.

Another place where Google has the advantage is full-text indexing. So far, OAI tools only search metadata. The very welcome OAI reply is that full-text indexing is coming. For one approach to it, see the work on the OA-X protocol.

http://web.archive.org/web/20040411030821/http://eepi.ubib.eur.nl/iliit/archives/000471.html

Sub-total

For every OAI virtue, there is some Google counterpart. This doesn't mean that the Google counterparts are superior or even equivalent. That will depend on variables such as your search skill, your search goal, and the year (remember, what we're calling Google is always improving).

I know you want me to choose between them but I'm not going to do it. If their merits really depend on your needs and circumstances, however, then this is already a kind of victory for Google, at least insofar as it means that putting an eprint on your personal web site won't *always* be worse, or won't be *much* worse, than depositing it in an OA-OAI archive. (If you're sorry that I'm not choosing between them, then here's a clue to my personal position: It was very difficult to bring myself to write out the previous sentence.)

Note how we have confirmed the wisdom of a general practice within the OA movement. If we provide OA to our eprints, then services to index and preserve them will come along after the fact. Depositing eprints in OAI-compliant archives makes those eprints fodder for all future OAI-compliant data services. Depositing eprints on a personal web site makes them fodder for all future iterations and rivals of Google. We don't have to wait for these services to emerge, or to reach a certain level of adequacy, before we provide OA to our eprints. On the contrary, we should provide OA to our work right now and let evolving data services compete to improve upon the visibility and longevity of our work for the rest of time. […]

But this is key: OA-OAI archiving and Google indexing are completely compatible. We can do both, and we should. That's the main reason why I'm not going to choose between them. […]