It's not filter failure, it's a discovery deficit

Based on presentations given at the RLUK 2010 Conference and STM Innovation Seminar 2010

The web has changed our information-seeking behaviour radically, yet scholarly communication remains firmly embedded in the traditions of the print world. Here, I argue that the dropping costs of publication and distribution mean that effort and resource expended on preventing publication is wasted and that developing the tools and culture for post-publication annotation, curation and ranking is more productive. Rather than see this as information overload, or in Clay Shirky's words, a 'filter failure', I propose that it is more useful to see the problem as a 'discovery deficit'. This flood of content, instead of being a problem, is an opportunity to build technical and cultural frameworks that will enable us to extract more value from the outputs of research by exploiting the efficiencies that web-based systems can provide.

CAMERON NEYLON
Senior Scientist, Biomolecular Sciences
Science and Technology Facilities Council
Rutherford Appleton Laboratory

The last generation of scientists to remember 'the library'

My first real exposure to research was in 1993. I vividly remember sitting in a laboratory while my research supervisor explained the process of research. At the heart of that memory is a phrase that nearly 20 years later sounds ludicrous:

> *"You need to spend half a day each week in the library. You should sit down and flip through all the new journals that have come in that week …"*

This was still a time when to find information you had to physically get out of the lab or the office and go to the library. The library was *the* place you went to find the information you needed, almost exclusively in journal articles.

This persisted through the beginning of my PhD in 1995, but as the world wide web gained a hold, journal articles started to appear in electronic format and tables of contents would arrive by e-mail. Medline, which had been nearly useless to me when it came on a compact disc, appeared online as PubMed, shifting the pattern of information discovery away from reading a paper index to running a search.

By the end of my PhD, in 1999, actually going to the library was a rare and arduous trek, only undertaken when I had built up a sufficiently large list of papers that were not available online. Today, another ten years on, I can't remember the last time I went into a library for something other than a cup of coffee. But this does not make the library less important to my work, just less visible.

The generation of scientific and engineering researchers that came after me live entirely in this new world. For them, information is obtained online. The library is a place for the occasionally necessary physical object, a place to study in, to meet with friends. When they can't obtain a text or a journal article due to lack of access it may not even occur to them to blame 'the library' for this omission. They will simply widen their search till they find what they are looking for (via legitimate routes or otherwise). Their 'library' is the whole of the internet.

The web changes everything, yet nothing has changed

The web and the systems that it has enabled have radically changed our ability to obtain, store and manage information. However, the forms in which researchers obtain that information have barely

changed at all. Journal articles, and monographs in the humanities, remain the premier form of scholarly communication despite obvious inefficiencies.

Cultural change is always slower than technological change, and the use of journal articles as a measure of researcher prestige is deeply ingrained in the research community. In a print-based world this made perfect sense, since the largest direct cost in the communication process was that of printing and distribution. The decision whether to publish an article or a book was therefore the key moment of community investment in that communication. Publishers decided to commit resources to production, and librarians subsequently decided whether to invest in the produced work. Both publishers and librarians took the role of gatekeepers, a crucial role that added value through filtering out what did not deserve to be published. The costs forced these decisions; not everything could be published.

But the world has changed. The human costs of the traditional publication process remain high. However, the costs of making information publicly available and distributing it have dropped to nearly zero. The explosion of online content, from Wikipedia to Lolcats, demonstrates that the marginal cost of publication *per se* is low enough that these services can be provided free to end users. Investing time and effort in explicit decisions about whether to publish no longer adds value, because the cost of publication is effectively zero and the potential costs of not publishing become significant. Curating and managing the wealth of information that is now available is where the value lies. It is also where the cost lies, and significant investment in terms of time and money is still required.

Information overload or filter failure?

The number of journals and journal articles being published is increasing exponentially both in response to the increasing size of the research community and the dropping costs of distribution. This increase, along with the increasing quantities of data being made publicly available, is leading to professional researchers finding themselves unable to cope: a classic case of information overload. So if we follow the above argument to its logical conclusion and publish everything, surely this will just make a bad situation worse?

This debate is often summed up using the soundbite which forms the title of a talk by Clay Shirky: 'It's Not Information Overload. It's Filter Failure'.[1] Shirky argues that we have always suffered from information overload, that the arguments advanced against the quantity of information are the same as those advanced every time a technological advance in communication occurs: from writing and literacy to today's online technologies. Each time, we adapt to these technological advances by building new filters for ourselves to find the content that we want.

The soundbite in and of itself does not do justice to the details of Shirky's argument. In the context of scholarly communication I believe it is particularly unhelpful, because it encourages us to support the status quo rather than to take advantage of the opportunities of the web. If we focus on filters, and particularly on the role of publishers and collectors as filters, then we will fail to exploit the real opportunities that Shirky identifies. In particular, the network effects that occur when a community can *collectively* mark up and comment on published work can make the process of discovering the works that are important and valuable to a particular user much more efficient. If researchers as a community fail to take advantage of these efficiencies, if we continue to encourage publishers and librarians to act in the role of gatekeepers, if we don't start taking responsibility for our own filters, then we are failing in our duty to maximize the efficient use of public money invested in research.

Let us consider a traditional research article, perhaps one that is somewhat controversial and has sparked off much comment both in the 'new' and 'traditional' media (recognizing that the distinction is increasingly meaningless). All of this commentary can add value to the article. In particular, it can increase the opportunities for others to discover and extract value from that article, even if the primary findings the article reports are wrong.

The decision to publish is a binary one, and one that comes to us from the print world. The article is either between a specific set of printed covers or it is not. The marking of an article as approved and its publication are inextricably linked. That this link should be broken, is compellingly demonstrated by retracted articles. They are still there, still 'part of the record', but they are somehow no longer 'approved'. On the web it is at least possible to

mark up a paper as 'no longer approved' while leaving it in place.

But this is still binary: published or not; retracted or not. Imagine a different world, in which the findings might be marked as incorrect but the actual data as valid – the methodology as validated and useful to some researchers, even if the data was poorly analyzed. The potential value of research lies in its potential to influence future research, to improve its outcomes, and to hasten the application of those outcomes. If I ought to know about the methodology or the raw data, if these could improve my current or future research, enable me to improve patient outcomes, or create an economic benefit, then delaying publication is actually destroying significant potential value to me and to the wider community. If the delay is because reviewers and editors can't agree on the conclusions of the paper then that disagreement is probably not even adding any value. The discussion of conclusions and whether the data support them is almost certainly better carried out by the relevant research community as a whole and not by a small number of unnamed reviewers. A single centralized binary filter can, and does, limit value.

It's not filter failure, it's a discovery deficit

The problem with focusing on filtering, and in particular on centralized filtering, is that different people need different filters at different times. The filter I need as a person exploring a new research area is different from the one I need when I want to check the details of how (or whether) to carry out a specific experiment. The filter I need when wanting to ensure that my introductory text covers all relevant previous work is different from the filter I want to use to decide what microscope to buy.

When we talk about filtering in the context of scholarly communication we are almost always talking about centralized mechanisms that block publication or block access. What is much more productive is to think of the problem not as 'filter failure' but as a 'discovery deficit'.[2] We lack the tools to reliably find what we need when we are looking for it. It is difficult for us to see clearly from our current standpoint what these tools will look like. But there will be tools as effective and as groundbreaking as a card catalogue or a book index, both tools that were unnecessary, even unthinkable, prior to the development of the technologies that made them feasible and the technologies that made them absolutely necessary.

Some of these tools are already with us and the ones that are popular focus on enabling discovery across the widest possible range of content. Think of your own information-seeking behaviour. When was the last time you used an encylopedia or a curated reference source? Most of us overwhelmingly use Google to obtain information. Google, and other search engines, enable discovery not across a limited subset of information but across the whole of the web. Search engines are a filter, but one that gives the user significant control, and they act to make it possible to discover content from anywhere on the open web.

The web is full of things we do not want to see but which can still actually benefit us. This material can act as false positives to improve discovery algorithms so that you find what you are actually looking for. If we stop viewing the flood of information as a liability, instead regarding it as valuable data to design and test better discovery systems, then we can extract real value while simultaneously preserving the potential that even a single person might find value in a given article.

Every book its reader

These two coupled ideas, 'publish everything', and 'focus on enabling discovery', shift our thinking about publisher and librarian roles both subtly and radically. It is not a massive step to move from thinking about these roles as selecting on behalf of readers, to thinking of them as guiding readers to the right content. Both publishers and librarians already provide tools that support discovery. It is just a shift in emphasis from selection to guidance.

However, it is a much more radical shift to follow this through, to realize that as long as that guidance is effective, publishing something can do no harm. To make the guidance effective we will need to retain and build on the traditional tools and approaches we have used to mark up the quality of published materials and to build systems that enable us to apply these quality marks across the web of content, but we need to abandon the notion that publication in and of itself is a quality mark. Breaking this link offers the greatest single opportunity available to us to realize more value from our research outputs. Peer review will not go away; it will simply be

separated from publication, enabling us to utilize it more efficiently, applying it to those publications that are worth the effort. A significant proportion of papers are never cited[3,4]; should they really be peer reviewed?

If people are critiquing a paper openly on the web they are adding value in the form of annotation and opinion. The quality of that annotation will vary greatly but there is potential value in this discussion that should ideally be made accessible from the paper. This discussion, this annotation, this curation of content will ultimately be of value in enabling discovery as well as rating and approving. A key part of discovery is to rank and prioritize potential hits. Discussion of research is implicitly a ranking process. But there are many ways that research outputs might be measured, and different user needs will require different measurements. The choice of how to prioritize research of interest should be in the hands of the user, not centralized in a place remote from them.

For those from a library science background, this is in many ways a return to the core values espoused in Ranganathan's Laws[5]. *'Books'* (in the most general sense) *'are for use'*, not for locking away because someone might misunderstand them. *'Every reader their book'*, the role of the curator and the librarian being to guide them to it. But, above all, it is the third law that is most apposite to the world we live in: *'Every book its reader'*. Every book, every piece of content has a value, a potential that can only be realized if it can be placed into the hands of the right reader. The right reader might be a researcher, a patient, or a student. Or the right reader might be a machine that will use the content to build something more, automating an annotation process, connecting the work with other relevant research outputs, or checking the consistency of conclusions with other work. What is important is that every reader finds their book, and every book and article and dataset, its reader. For the first time in history we live in a technological environment where this is, if not yet straightforward, then technically conceivable.

The gatekeeper is dead! Long live the gatekeeper!

This vision may be compelling but it is just that, a vision. What can we say in concrete terms about how this world can be brought about, and what does it mean for publishers and librarians charged with the management of scholarly communications? The role of the gatekeeper is no longer one that adds any value. If your work involves restricting access, preventing publication, even selecting content for purchase, these roles will disappear over time. The flip side of these roles, enabling access, providing the infrastructure for publishing and services that can add value, building the technical frameworks for curation and ranking, will continue. Above all, enabling discovery and the role of 'discovery experts' will grow. The central gatekeeper is dead. We are now all our own gatekeeper and we will all need help.

In terms of mechanisms, it is valuable to examine the web services that support a wide range of communities with different needs to collaboratively discover, annotate and rank web-based content. From Facebook to Twitter, StackOverflow to Quora, relevant content is being created, identified, ranked and shared within communities. These systems, crude as they are, systematically offer more options for searching and filtering to the end user than our traditional publication system. There are large opportunities here to apply the technical expertise within publishing and library organizations but there are also significant social and cultural challenges.

A significant challenge is how to transfer, both in technical and in cultural terms, the perceived and real value of traditional peer-review processes into a web-native world. There are massive cultural barriers here. Even the idea that something 'wrong' should be published is anathema. But these barriers can be balanced against the huge need for better discovery mechanisms and the limited resources available for annotation and curation. Something has to shift but it is likely that incumbent information providers, both publishers and indexers, will have a role to play here.

The library is no longer a building but is the whole world. Communities of researchers, communities of students and of the interested public will need expert assistance to find what they are looking for. This assistance might be direct, person to person, or it might be mediated through technology. Equally, the sheer volume of information that needs to be curated and annotated means that a centralized professional community cannot cope. You will need to enable us to help you do this mark up through providing and managing the

frameworks that aggregate and integrate this information. Either way, systems need building, researchers need training. The traditional roles of curation, archiving, indexing, discovery, all remain. Even selection remains, but the selection is not for publication, but for expending the limited resources available for annotation and curation.

The problem is not that we have too much information. We are an information-driven society; how could there be too much? The challenge is to make effective use of it. We don't need to block. We don't need to limit. We need to enable. We need the tools for discovery. This is not a problem. It is an opportunity, and we will make much faster progress in solving the problems we face when we see it for the opportunity it is.

References

1. Web 2.0 Expo NY: Clay Shirky (shirky.com) It's Not Information Overload. It's Filter Failure: http://web2expo.blip.tv/file/1277460/ (accessed 6 January 2010).
2. Neylon, C, It's not information overload, nor is it filter failure. It's a discovery deficit, *Science in the Open*: http://cameronneylon.net/blog/it's-not-information-overload-nor-is-it-filter-failure-it's-a-discovery-deficit/ (accessed 6 January 2010).
3. Minjers, J and Xu, F, The Drivers of Citations in Management Science Journals, *Kent Business School Working Paper Series*, 2009, 198: https://readinglists.kent.ac.uk/kbs/documents/research/working-papers/2009/198-regression_paper.pdf (accessed 6 January 2010).
4. Weale, A R, Bailey, M and Lear, P A, The level of non-citation of articles within a journal as a measure of quality: a comparison to the impact factor, *BMC Medical Research Methodology*, 2004, 4:14.
5. Wikipedia: Five laws of library science: http://en.wikipedia.org/w/index.php?title=Five_laws_of_library_science&oldid=405924255 (accessed 6 January 2010).

Acknowledgements

The author would like to thank Christina Pikas and Dorothea Salo for their helpful comments on the first draft of this paper.

■ **Cameron Neylon**
Senior Scientist, Biomolecular Sciences
Science and Technology Facilities Council
Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot OX11 0QX, UK
E-mail: c.neylon@rl.ac.uk

This theme will further be explored by Cameron Neylon in his presentation 'The gatekeeper is dead! Long live the gatekeeper! *Or*: What does filtering mean for scholarly communications in a web-based world?' to be given at the UKSG Conference on 4 April 2011 in Harrogate.