nature webdebates

Distributed and centralized technologies: complementary tools to build a permanent digital archive

Matt Cockerill
Technical Director, BioMed Central Limited
matt@biomedcentral.com

In the debate on open access to the biomedical literature, one of the arguments used by opponents of PubMed Central has been the suggestion that central archives and indices are technologically unfashionable. For example, Ira Mellman claims that "it is outmoded and incorrect to contend that hosting content on single sites is necessary for complete searching."

It would be a great shame if the hype surrounding "distributed" technologies (often known as "peer-to-peer" or P2P) were to obscure the essential role played by central archives and indices, and the way in which centralized and distributed systems complement each other.

Central indices facilitate efficient searching and mining of data.
Existing Web search engines index relatively unstructured data – hypertext markup language (HTML) pages. But by bringing this data into one place to index it, a centralized Web search engine such as Google is able to extract information concerning the data as a whole (for example, statistics on linking between sites) and use this to improve the quality of the search results delivered. Using its central index, Google is also able to generate a list of the most closely related pages for any given Web page. When indexing highly structured data such as biomedical research articles, there is even more opportunity to extract such valuable information about the full corpus, while building the central index.

Although it is certainly possible to send search queries to multiple search engines and then combine the results returned (as is done by Metacrawler and other Web metasearch engines), there are fundamental technical obstacles to scaling this approach to integrate large numbers of indices held in different locations. With many servers to be queried, even the simplest search would take unacceptably long, consume excessive amounts of bandwidth and deliver erratic results, depending on the availability of the various servers.

To take a simple example, if a user wants to find the 100 most recent articles mentioning a particular disease in a centrally indexed database such as PubMed Central, the query to do so can be processed very quickly and efficiently. But to do the same query across a network of 1,000 individual journals' sites would be extremely inefficient. It is not possible to know in advance which servers will contain the most recent articles so, to find them, data from all servers must be compared, directly or indirectly. Sophisticated algorithms can go some way to making this more efficient – and this is an active area of research (see for example http://www.jxta.org) – but performance, bandwidth and scaling issues mean that distributed search engines are not an attractive choice for querying the scientific research archive for the foreseeable future.

Fortunately, the growth of the Web has driven the rapid development of search-engine technology, which can now be used to index and query terabytes of text data at reasonable cost[1]. The continuing decrease in the cost of server hardware means that it is both technically and economically feasible for public or privately funded repositories to be developed which would index the full corpus of published scientific research and allow it to be searched rapidly and efficiently.

Making content accessible from multiple independent repositories increases reliability, guarantees permanence and encourages innovation in content delivery.
Recently, as Sequeira *et al.* have said in their contribution to this debate, PubMed Central has agreed to allow publishers to participate by submitting data to be indexed and archived at PubMed Central, while ensuring that requests to view the full text of any article would be forwarded to the publisher's website. Although this is a pragmatic stepping stone towards open-access to research, it would be far better from the point of view of the scientist if a copy of the research were made available for download directly by independent digital archives (including, but not limited to, PubMed Central).

Anyone who has used the Web extensively will be aware of the practical reality that any particular Web server will not be accessible 100% of the time. There are many reasons a user may be unable to access an article at a particular site, including: network connectivity issues at any point between the site and the user; broken links, because of reorganization of the site concerned; software or hardware problems affecting the site concerned; and servers being overloaded with traffic at popular times (such as on the day a new issue of the journal appears). This is why Google makes available a cached copy of all the pages that it indexes. Similarly, Research Index, NEC's huge index of computer-science research articles, not only links to the articles on the sites where they were found but also offers a cached copy for download.

As well as providing reliable access in the short term, storing copies of research in many different archives is also the most obvious safeguard to assure the permanent accessibility of research in digital form, independent of the survival of any particular organization or publisher.

Perhaps most importantly, if multiple central repositories are able to deliver research, rather than being forced to refer browsers to the publisher site, then this will enable each repository to offer a simple coherent interface to all the scientific research it indexes. It will also foster innovation in the creation of front-end tools to deliver, interlink and annotate research articles in the most effective way possible.

Distributed technologies and protocols enable automated exchange of data between sites.
Critics of PubMed Central have also suggested that for publishers to make their data available for archiving and indexing at PubMed Central would be both expensive and error-prone. For example, Ira Mellman says: "Posting this information on a second site is fraught with – difficulties – No software package or established standard yet exists that can guarantee the accuracy of each re-posting. Catching errors will be at best difficult."

Fortunately, automated data exchange is an area where much technological progress has been made in the past five years. XML provides a simple yet powerful framework for structuring data such as scientific research articles and exchanging them with other organizations. Specific technologies such as LOCKSS and the Open Archives protocol, which is already supported by BioMed Central, have been developed to allow data to be automatically exchanged and updated between article repositories without human intervention. More generally, developments such as the Simple Object Access Protocol, an emerging standard for XML-based server-to-server communication, are helping to make this type of data exchange even more reliable and cost-effective.

If all published research were to be distributed freely through such systems, then any number of organizations could collect data and build independent archives and indices. Some of these archives might aim to be comprehensive by collecting and indexing all available data. Others might instead

focus on particular subject areas. Each archive would be free to develop its own tools to add value to the data and to mine it for useful information. This model offers far more power and flexibility to scientists than the current model, whereby each research archive forms an island of content that cannot easily be integrated with the content from another publisher.

The introduction of errors when converting between data formats clearly must be guarded against, but the use of XML makes this relatively straightforward. In fact, as noted by Sequeira *et al.*, when a publisher shares its data with other organizations this generally has a positive effect on the quality and integrity of that data. When a repository receives XML data, it will interpret it on the basis of structure rules laid down in that XML's document type definition (DTD) or the DTD's more powerful successor, the XML Schema. If there are problems with the data's DTD or Schema compliance, these will be identified quickly during processing. This automated feedback will help publishers to improve the quality of their data. Without these checks, a publisher may not notice data problems and everything may appear fine on their website, until at some point in the future the website is changed or the data is reused for another purpose and the problem emerges.

Finally, this model of multiple communicating archives makes transparent what should already be obvious – that existence of one or more central indices and archives in no way implies a central point of control for what can be published, any more than the existence of search engines such as Google or AltaVista implies central control of what may be placed on any particular website.

Various technical arguments have been levied against PubMed Central but these do not stand up to scrutiny. Rather than arguing the relative merits of central and distributed technologies, forward-thinking publishers need to make intelligent use of both to deliver the best possible service to scientists.

Reference

1. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. Proc. 7th Int. World Wide Web Conf. (1998).

After completing a PhD in Biochemistry with the Imperial Cancer Research Fund, Matthew Cockerill joined Current Science Group, where he was closely involved in the development of BioMedNet, one of the first portals for biologists and medical researchers. In late 1999 he rejoined Current Science Group to develop BioMed Central, a new publisher which uses online tools to allow scientists to publish peer-reviewed research with immediate, free online access for all.