Traditional publisher experiments with Open Access

When an archetypal traditional publisher like Oxford University Press announces that it is to begin experimenting with Open Access publishing models, it is a clear indication that the tide is changing. The largest university press in the world announced in August that it will launch an Open Access initiative “in response to calls from the academic community to make research freely available online without the barrier of a subscription to access.”

Oxford University Press (OUP) has been around for centuries. “Oxford University Press had its origins in the information technology revolution of the late fifteenth century, which began with the invention of printing from movable type,” boasts its website. OUP is a ‘department’ of Oxford University, one of the world’s oldest scholarly institutions. As such it enjoys charitable status and tax exemption; OUP transfers 30% of its annual profits to the rest of the university, with a commitment to a minimum transfer of £9 million per annum.

OUP is to initiate its Open Access experiment with one of its flagship journals, Nucleic Acids Research (NAR). NAR has an impact factor of 7.05 and has been listed by ISI as one of the top ten ‘hottest’ journals of the decade in biology and biochemistry. NAR is to adopt an ‘author-funded’ Open Access model for the publication of its annual Database Issue to be published in January 2004.

This will be the eleventh issue of the popular database compendium that is compiled by Andy Baxevanis, director of the Computational Genomics Program at the National Human Genome Research Institute, Bethesda, MD. The papers in the 2004 issue will be freely available online from the moment they are published. The journal announced that it will gradually move to an Open Access model over a transition period of 4-5 years, at which point it expects all research published in NAR to be funded in this way.

“As a University Press we believe in making research as globally accessible as possible, but very few well-established journals have made a move to Open Access,” said Martin Richardson, director of the OUP Journals Division that publishes over 180 scholarly journals. He believes that the transition approach is important for authors and readers. “Unless funding conventions change, and all authors have access to sufficient financial resources to pay for the full publication costs of their research, it is unlikely that Open Access publishing would be widely adopted by well-established journals,” says Richardson.

Increasing numbers of high-profile journals appear willing to experiment with ‘author pays’ models. The British Medical Journal (BMJ) has been freely available online for almost a decade. During that time its impact factor has continued to rise, outperforming other similar journals. The editor, Richard Smith, said that BMJ will now start to introduce a minimal charge for online access to review material and will experiment with author charges for research articles.

Smith said that BMJ will continue to deposit original research articles in the PubMed Central archive. It is not yet clear whether NAR will also deposit in PubMed Central, a requirement for ‘full Open Access’ according to the Bethesda principles (allowing free redistribution of full-text articles; see Open Access Now, July 14, 2003 for a full definition).

In a statement, OUP said that it hopes that its NAR initiative will help lead the way in exploring alternatives to the subscription model that will make research literature as widely accessible as possible.

http://nar.oupjournals.org
http://bmj.com
FEATURE: Data mining

Data mining: Open Access research

By Matthew Cockerill, Technical Director, BioMed Central

The need for data mining tools

The biomedical research literature is growing at a daunting rate, with more than 400,000 new research articles listed each year in PubMed alone. The power of post-genomic research technologies and increases in biomedical research funding suggest that we can expect the number of articles to continue to expand at an overwhelming rate. The sheer scale of the scientific literature poses a formidable challenge to research scientists in their attempt to locate published results that are relevant to their research interests.

Text searching of abstracts and full-text articles has helped. But much information is missed by simple text searches. More sophisticated tools are needed to identify relationships, connections, and patterns that are hidden within the data reported in the scientific literature. Data mining researchers and computer scientists are developing such tools using statistical, linguistic and artificial intelligence approaches.

The search engines used to find information on the Internet provide a simple example of how such techniques can be helpful. The well-known search engine Google exploits the pattern of links between different websites in order to generate a 'page rank' for each page that it indexes. Pages that are frequently linked to by other pages are given a higher page rank. And it turns out that page rank is an extremely helpful predictor of how useful a page is to Internet surfers.

Data mining experts hope to develop tools that can identify knowledge that lurks beneath the surface

But new approaches are needed to tackle both the expanding Internet and the scientific literature. The World Wide Web was originally built for human consumption, and although everything on it is machine-readable, this data is not machine-understandable. Part of the problem comes from the very nature of natural language that presents a considerable challenge for computational algorithms to extract meaningful information. For example, English has both synonymy (multiple ways to refer to the same object) and polysemy (words with multiple meanings), which complicate text searching. Scientific literature has numerous examples of such ambiguities. For example, individual genes often have several different names that relate to their historical discovery. And names such as RAS and MYC are used to refer to both genes and their protein products.

More sophisticated approaches are required that can extract meaning taking into account the context within the text. Furthermore, data mining experts hope to develop tools that can identify knowledge that lurks beneath the surface and can be only found by connecting the 'semantics' (meaning) reported in one article with findings reported in another.

Semantics, mark-up languages and ontologies

The parallels between finding information on the Internet and finding relevant information in the scientific literature can offer instructive hints for data mining approaches.

Jim Hendler, a computer scientist at the University of Maryland, is a pioneer of the Semantic Web, a project developing Internet resources for the future. “The Semantic Web is not a separate web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation,” explained Hendler in a recent article. “For the Semantic Web to function, computers must have access to structured collections of information and sets of inferences rules that they can use to conduct automated reasoning,” says Hendler.

That requires a computational language that can be used to express both data and rules for reasoning about the data. The advent of extensible markup language (XML) as a standard for information interchange holds invaluable promise, as XML makes it possible to express mathematical equations and chemical structures in a way that a computer can easily recognize and process. XML gives structure to a document by creating tags, hidden labels attached to annotated information. Resource description frameworks (RDF) are then used to express meaning, which is encoded in sets of triples, each triple being like the subject, verb, and object of an elementary sentence.

The third key component of a functional Semantic Web is a system that allows algorithms to define the relationships between different terms – this is referred to as ontology. The use of ontologies is illustrated by the Gene Ontology (GO) Consortium that is developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner.

Once biological research information is structured in this way, using appropriate mark-up languages and ontologies, researchers will have a realistic opportunity to mine the data. This paves the way for computer-assisted discovery.

“The main benefits are realized by using these techniques to create curated machine-readable datasets,” says Ian Donaldson at Samuel Lunenfeld Research Institute in Toronto. “Biomolecular knowledge is really contained in the collective mind of the research community. The biomedical literature is an imperfect reflection of this knowledge that is obfuscated by natural language and its use in describing models of consensus that are constantly changing. Data mining techniques facilitate one step in creating

In the words of Sydney Brenner

The definition of data mining: what’s my data is mine and what’s yours is also mine.
Donaldson has developed PreBIND, a data mining tool that helps researchers locate biomolecular protein-protein interaction information in the scientific literature. Users can enter the name or accession number of a protein and PreBIND will return a list of potentially interacting proteins. The list of potential interacting proteins is determined using a list of protein synonyms, in combination with a supervised learning algorithm known as a support vector machine (SVM). The SVM determines whether there is an indication in the machine (SVM) that there is an indication in the literature that would be missed by a simple search. Donaldson says that there is no doubt that Open Access to full-text articles is essential if the promise of data mining is to be fully realised. Only then will data mining lead to the discovery of new horizons. “The human mind has an attention horizon. While you’re able to go anywhere in the world, you’re unlikely to visit that place without knowing of its existence and how it relates to something near you. Data mining expands the circumference of that horizon.”

BioMed Central is leading the way in encouraging researchers to do data mining research, by making available its entire corpus of Open Access research available for download via ftp in XML form (see www.biomedcentral.com/info/about/datamining). We hope other publishers will follow.

Glossary of data mining

Data mining
Using computers to extract information that is hidden within a large set of data. Data mining researchers who work with text use statistical, linguistic, and artificial intelligence techniques to go beyond simple text searching.

BioNLP.org – natural language processing of biological text. www.ccs.neu.edu/home/futrelle/bionlp
BioMed Central’s data mining information page www.biomedcentral.com/info/about/datamining

Semantic Web
The vision of Tim Berners-Lee for an improved version of the World Wide Web in which content is annotated in a machine readable way, to allow its meaning to be analysed by automated ‘agents’. www.semanticweb.org
www.mindswap.org/Science

XML
eXtensible Markup Language (XML) is a standard text format that allows information to be represented in a structured way, thereby facilitating automatic processing. Different dialects of XML (known as schemas) are used to describe different types of content (for example, CML describes chemical structures, MathML describes equations). www.w3.org/XML
www.xml-cml.org
www.w3.org/Math

Ontology
In artificial intelligence research, an Ontology refers to a structured collection of concepts relevant to a particular domain of knowledge. For example, an Ontology might incorporate the concept of engrafted, which would be an instance of the concept gene. And like all genes, engrafted would be associated with a specific organism (Drosophila), and chromosome (2).

Gene Ontology Consortium www.geneontology.org

PreBIND
PreBIND is a data mining tool that helps researchers locate biomolecular protein-protein interaction information in the scientific literature.

PreBIND home page – www.bind.ca/index2.phtml?site=prebind
Article on PreBIND – www.biomedcentral.com/1471-2105/4/11
Open Access Now has attracted the attention of many people involved in scholarly publishing. We have received letters and emails from a wide spectrum of readers including research scientists from around the globe, Open Access advocates, librarians and members of the public. The Open Access Now website includes a Letters page that features selected letters from readers, and provides the opportunity to comment on letters and articles. The debate about Open Access tends to be dominated by a fairly small number of well-oiled voices. Open Access Now provides a forum for the many scientists who have not yet aired their opinions about these important issues. There are increasing signs that if readers and authors make their views felt, traditional publishers will bow to the pressure. Now is the time to share your concerns and participate in the debate. Open Access Now can give a voice to the silent majority – so make yourself heard!

BOAI was launched in February 2002 as a result of a meeting in Budapest organised by the Open Society Institute. The meeting attendees co-signed a declaration that now invites signatures from individuals and organisations who wish to pledge their support. It asks signatories – including editors, publishers, librarians, researchers and many others – to join the effort to ensure a transition to Open Access publishing.

Why does BOAI exist?
The aim of the Budapest meeting was to explore how the various ongoing initiatives to free research from the constraints of traditional publishing could work together, and how OSI and other organisations could most effectively and affordably use their resources to ensure the success of the Open Access movement. Now, BOAI invites “the signatures, support, and participation of the entire world scientific and scholarly community.”

www.soros.org/openaccess

What do you think you gained from publishing in an Open Access journal?
Rapid acceptance-to-publication time was very important for this paper. BioMed Central provided an extremely efficient submission process and allowed my published work to be viewed without restriction by the rest of the scientific community. I have since recommended publication with BioMed Central to several colleagues who expressed interest after seeing my published article. Each of them was as pleased as I was with the excellent service provided.

An author speaks...

Dennis Maddox
Medical College of Georgia

Dennis Maddox is a postdoctoral fellow in the Department of Cellular Biology and Anatomy, Medical College of Georgia, USA. He was co-author of the first research article published in BMC Developmental Biology, and has recently published another article in the same journal.

What prompted you to submit your first article?
My advisor and I submitted the very first manuscript to BMC Developmental Biology. We submitted our manuscript to BioMed Central because we were excited about the potential of Open Access publishing and the amount of exposure that our research could receive when our findings could be freely accessed.

How would you describe your experiences of publishing with BioMed Central?
One major factor in choosing to submit our manuscript to BMC Developmental Biology was the promise of simple submission practices and rapid turn around time. In my experience, BioMed Central provided an extremely efficient submission process and allowed my published work to be viewed without restriction by the rest of the scientific community.

Research news from BioMed Central journals

Health-food-store advice could be bad for your health
Patients with breast cancer may be at risk if they follow misleading advice from health-food-store employees. Research published in Breast Cancer Research shows that store assistants often recommend products, with no proven benefit, that are expensive and potentially harmful. The Canadian study reports “health food stores are recommending a variety of products, [...] of which are supported by sufficient evidence of efficacy.” The average cost of these products was $58 per month, with more expensive remedies up to $600 relying on “insufficient or questionable research”.

Furthermore, many recommendations lacked discussions of possible adverse effects of the product, or potential interactions with conventional cancer treatments. Breast Cancer Research 2003, 5:R170-R174.

Why are elephants bigger than mice?
Research published in Journal of Biology uncovers a key pathway that controls the number of cells in an animal, thereby controlling its size. Ernst Hafen and colleagues, from University of Zürich, investigated the role of the Drosophila Forkhead-family transcription factor, dFOXO. The authors used a combination of genetic techniques to demonstrate that the dFOXO gene mediates the reduction in cell number that is one of the effects of reduced insulin signaling. This is partly due to dFOXO’s ability to regulate expression of an inhibitor of translation initiation, 4E-BP.

Flies with no functional dFOXO looked normal but were more sensitive to oxidative stress. Journal of Biology 2003, 2:20.