The medical journals back-files digitization project and open access

Based on a paper presented at the 29th UKSG Conference, Warwick, April 2006

This article discusses the medical journals back-files digitization project – an initiative to create a critical mass of digital content based on the back-files of a number of historically significant medical journals. All content is made freely available through the NIH life sciences repository, PubMed Central (PMC).

The project, however, is not limited to historical, archival content. All publishers that participate in this initiative agree to deposit current and future issues with the PMC archive, and make all research papers freely available within an embargo-defined period. As such, this project is in accord with the open access principles espoused by the project sponsors – the Wellcome Trust, the JISC, and the US National Library of Medicine.

This paper considers some of the value-added services this project has developed, and explores why the Wellcome Trust is supporting open and unrestricted access to the research literature and what it is doing to help realize its open access policy.

Back-files project: an overview

The medical journals back-files digitization project is an initiative to create a free-to-access, critical mass of digital content, based on the back-files of a number of historically significant medical journals.

Carried out in partnership between the Wellcome Trust, the Joint Information Systems Committee (JISC), the National Library of Medicine (NLM) and a number of publishers, this project will, on completion, deliver more than 3 million pages of content, made freely available on the Internet via PubMed Central (PMC) http://www.pubmedcentral.gov.

From the start, however, the project sponsors wanted a dynamic, live archive that would continue to grow. Consequently, in addition to digitizing back-file content, participating publishers agree to deposit all current and future issues of the relevant titles with PMC.

Thus far, 17 journals, published by a mix of for-profit, not-for-profit and learned society publishers, have agreed to participate in this project. Titles that will be made available through this project include the Annals of Surgery, BMJ, Biochemical Journal, and the Journal of Physiology. A full list of titles that are participating in this project can be found at http://library.wellcome.ac.uk/backfiles.

To date (May 2006), around 1 million pages have been digitized and are online. The digitization programme will continue over the next 18 months (and possibly beyond if additional titles come on board), with content being released in PMC as it becomes available.

A value-added product

In addition to creating page-scans (TIFF images) of every single page in the archive, this project has sought to provide a number of value-add features that will help researchers exploit the full potential of the archive. Some of these key features are described below.

Searching the archive

Every word within the archive is indexed (using optical character recognition – OCR), thus enabling...
a researcher to look for any word or phrase, wherever it may appear within an article.

In addition, we also create a bibliographic citation for every discrete article (e.g., research paper, editorial, review, letter, etc.), for inclusion in PubMed. Thus, a researcher who uses the PubMed database to find articles relevant to the research they are undertaking will be alerted to articles (with links to the full-text article in PMC) irrespective of when they were published or when PubMed/Medline started indexing that title.

Finally, as Google (and other search engines) are able to crawl and index every paper within the archive, the researcher who uses these generic search tools will still be made aware of relevant papers within the back-file archive.

Innovative linking
To help the researcher move from one relevant research paper to another, all references cited in a paper are programmatically extracted from the TIFF image and rendered as hypertext links. Thus, if a paper references another study that is included in the PMC archive, the researcher can simply follow the link and view the article (Figure 1).

This linking process is run dynamically, so as more content becomes available in PMC the chance of being able to move from article to article is increased. If the article is not available in PMC, the reference will be linked to a PubMed citation (if one exists) and from there – dependent upon the publisher’s policy – it may be possible to link to the original paper.

In much the same way it is also possible to identify a single paper and determine which other papers in the archive have cited that work. Figure 2 shows that more than 1,400 other papers in the PMC archive have cited Burton’s classic 1956 paper on assaying DNA1.

Another powerful linking facility relates to online corrections. When a journal within the archive includes a published erratum, this data is tagged in the PMC database and a link is created to the original article (Figure 3). Providing this level of functionality in the analogue world is something few librarians would be able to achieve!

Figure 1. Linked references: the researcher has the ability to link seamlessly from a citation to either the full text (if available in PMC) or to the PubMed citation.
Figure 2. The researcher has the ability to find other articles in the PMC archive that have cited a specific work.

Figure 3. Corrections are linked back to the original paper.
Linking research papers to their underlying datasets is one other innovative feature that has been developed within PMC. Figure 4 shows, for example, how a paper is linked to related datasets, such as chemical compound and gene sequence databases. As this linking is two way, the researcher can either move from the text to the data, or vice versa.

**Faithful replication**

In digitizing the back-files we strive to produce a completely faithful replication of the published journal. Thus, as well as digitizing all the papers within an issue, we also digitize any advertisements, journal administrative matter, tables of contents and covers.

To ensure that the archive is an accurate representation of the paper version, an extensive quality assurance programme has been developed by the NLM. This involves taking a sample of the archive (up to 10%) and comparing it with the paper version. If any pages are missing, or images appear skewed, etc., then the batch is rejected and the contractor is required to re-scan.

Further, every photograph that is found within an article is subject to a separate high-quality image.
scan, which is also made available to the reader (Figure 5).

**Publisher participation**

The success of this project is highly dependent upon the support from publishers. Without their content, there is no archive.

Publishers who agree to participate in this project not only allow their archives to be digitized (and be made freely available) but also agree to deposit all future issues in PMC.

Current content can be embargoed for an agreed time. In brief, research papers must be made freely available, through PMC, within 12 months of publication, whilst all other content must be freely accessible within 36 months.

Publishers who wish to host a copy of the backfile archive on their own sites (in addition to it being made available in PMC) can do so, and they are free to re-package (and sell if they so choose) this content in any way.

**Open access to research outputs**

This project provides a unique opportunity to open up access to medical history as recorded in the pages of historically significant medical journals. It demonstrates the value of ‘joined up’ partnership projects, and it fits in with the Wellcome Trust’s Strategic Plan, in terms of providing resources that contribute to a long-term and vibrant research environment. It is also in accord with the Trust’s position on the desirability of open access to the scientific literature. For the remainder of this article, I will focus on the issue of open access and the Trust’s role in taking this agenda forward.

**OA policy of the Trust**

The Trust was the first research funder to introduce a grant condition that requires grantees to make their research papers freely available to all.

Specifically, Trust grantees are required to deposit in PMC electronic copies of any research
papers that have been accepted for publication in a peer-reviewed journal and are supported in whole or in part by Wellcome Trust funding. All deposited papers must be available, for free, no later than six months after the official journal publication date.

Recognizing the significant change this introduced to research publishing, the Trust phased in this condition over a 12-month period. It applied to all new papers arising from grants awarded after 1 October 2005 but does not come into force for existing grants until 1 October 2006.

In addition, the Trust has provided grant holders with additional funding to cover the costs of page processing charges levied by publishers who support the open access (or hybrid OA) model.

**The driver for OA: a funder’s perspective**

For a funder like the Wellcome Trust, providing open access to the literature it has funded is attractive for a number of reasons.

Firstly, it is a fundamental part of our charitable mission to ensure that the work we fund can be read and utilized by the widest possible audience. Unrestricted distribution, via the Internet, currently offers the best available method to do this. In contrast, the current ‘reader/librarian-pays’ model significantly restricts access. For example, a survey undertaken by BioMed Central found that fewer than half of the articles resulting from NHS research grants are accessible online to NHS employees.

Secondly, providing open access to the research literature enables these outputs to be linked and integrated with other resources. Research papers that are tagged in a standard, uniform way – e.g. in compliance with the NLM Archiving DTD – can be read and searched by computers (as well as people), thus enabling context-sensitive links to be made to other online resources.

Finally, by mandating its grantees to make all research outputs accessible through PMC/UKPMC (see later), the Trust is helping to ensure that the digital record of medicine can be preserved. All papers that are added to the PMC/UKPMC repository are marked-up according to the NLM XML Journal Archiving DTD (http://dtd.nlm.nih.gov/). Mapping documents to this standard, non-proprietary format should ensure that future generations will be able to read these digital files, irrespective of developments in either hardware or software environments.

**Cost of open access**

The effect of the ‘author/funder-pays’ model is that it realigns the market, making publishing a research cost rather than a library cost.

Though the OA model provides free access to the literature for the reader, there are costs associated with this approach. For example, managing the peer-review process and copy-editing the final manuscripts are value-added services that incur costs.

Looking at the charges levied by both open access publishers and those publishers that have introduced a hybrid OA model (see Table 1), these, from the perspective of the Wellcome Trust, are affordable.

In a typical year the Trust is acknowledged in approximately 4,000 original research papers. If every single one of those papers was published as an open access article, and taking the average cost (£1,210 per article), the total cost to the Trust would be £4.84 million, or just over 1% of our annual research budget.

It is also worth noting that the Trust is rarely the sole funder of a research team. Indeed, more than 80% of papers that acknowledge our support also

<table>
<thead>
<tr>
<th>Publisher</th>
<th>Example title</th>
<th>Cost per article (assumes no discount)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioMed Central</td>
<td>Arthritis Research and Therapy</td>
<td>£750</td>
</tr>
<tr>
<td>Blackwell – Online Open</td>
<td>Journal of Physiology</td>
<td>£1,250</td>
</tr>
<tr>
<td>OUP – Oxford Open</td>
<td>Rheumatology</td>
<td>£1,500</td>
</tr>
<tr>
<td>Public Library of Science</td>
<td>PLoS Medicine</td>
<td>$1,500 (approx. £850)</td>
</tr>
<tr>
<td>Springer – Open Choice</td>
<td>Journal of Human Genetics</td>
<td>$3,000 (approx. £1,700)</td>
</tr>
<tr>
<td><strong>Average cost per article (across these 5 publishers)</strong></td>
<td></td>
<td><strong>£1,210</strong></td>
</tr>
</tbody>
</table>

*Table 1. Typical costs (April 2006) of publishing in an open access journal*
acknowledge the support of one or more other funders. In time, these costs will be spread throughout the research budget and will, we believe, fall below the 1% figure estimated here.

Much of the debate around OA has focused on traditional publishers and whether this model will bring about their decline. However, though the rhetoric has, at times, been apocalyptic in tone, what evidence there is suggests that open access publishing can provide a means by which publishers can continue to meet costs and turn a profit. Springer, for example, a traditional ‘for-profit’ publisher, believes that it can generate a profit through open access.

Further, a study that looked at the viability of the open access model for learned societies concluded that these publishers could continue to deliver the average surplus to their societies by introducing an OA fee of £1,166 per article (2004 costs).

**UK PubMed Central (UKPMC)**

The Wellcome Trust in partnership with a number of major UK biomedical research funders – the UKPMC Implementation Group – is working to establish a UK version of PubMed Central (UKPMC). The aim of this initiative is to create a stable, permanent and free-to-access online digital archive of the full-text, peer-reviewed research publications (and datasets) that arise from the research funded by the UKPMC Implementation Group.

Initially this service will mirror the data held in PMC, but longer term the objective is to create an independent resource that can offer enhanced linking and searching capabilities and be configured to meet the specific requirements of both UK researchers and funders.

UKPMC will also provide a manuscript submission, tracking and XML-conversion service, thus allowing researchers to ‘self-archive’ papers which otherwise would not be included in PMC (or UKPMC). To be clear, however, the Wellcome Trust would always prefer the final, published version to be available in PMC/UKPMC and, as discussed previously, has made additional funding available to meet the costs associated with providing this service. We anticipate UKPMC to be available early in 2007.

**Conclusion**

The Wellcome Trust believes that publication costs should be seen as a research cost – and as such should be borne by the funder.

In the near future we believe that the combination of open access publishing and the establishment of open access repositories, like PMC and UKPMC, will change the way in which biomedical research is disseminated.

The back-files project and open access initiatives, described in this paper, both provided unrestricted access to research outputs which, in turn, will help researchers to make the discoveries we need to improve human and animal health.

**References**

2. Details of the Trust’s policy on ‘Open and unrestricted access to research outputs’ can be found at http://www.wellcome.ac.uk/openaccess
4. Jan Velterop, Director of OA Publishing at Springer, commented: “we are absolutely convinced that with open access we can have good profit margins”. See *The Scientist*, 2005, Volume 6, Issue 1. Available online at http://www.the-scientist.com/news/20051011/02
6. Details of the UKPMC Implementation Group can be found at http://www.wellcome.ac.uk/assets/wtx028464.pdf
7. The mission of the Wellcome Trust is ‘to foster and promote research with the aim of improving human and animal health’. See http://www.wellcome.ac.uk/aboutus/whatwedo/