THE CRISIS IN SCHOLARLY PUBLISHING IN THE HUMANITIES

by John M. Unsworth, Associate Professor, Department of English, and Director, Institute for Advanced Technology in the Humanities, University of Virginia

Editor's note: In August, the author becomes Dean of the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. These remarks were presented at the 2003 Annual Meeting of the American Council of Learned Societies (ACLS) in a session on "Crisis and Opportunities: The Futures of Scholarly Publishing." The papers from the meeting are available on the ACLS Web site <http://www.acls.org/ex-03am.htm>.

In a paper given at Dartmouth College last November, and written in response to Professor Greenblatt's letter to the members of the Modern Language Association (MLA), I said:

If we can tackle large problems, with the resources of the computer, the network, and interdisciplinary collaboration, then stand-alone, single-author work on smaller problems is eventually going to seem ... quaint.

I realize that this is a tendentious statement, and I realize that my colleagues will regard its prediction as outlandish. But I do believe that the detailed and thoughtful fretting that we have been doing over the fate of the humanities monograph will seem—from the perspective of a not-too-distant future—beside the point, and all our carefully constructed responses to the crisis of scholarly publishing will be beside the point as well. In short, I believe we are, as the saying goes, preparing to fight the last war. I also believe that there is a way out of the present situation, although not without a leap of faith—or, to be more biblically precise, not without casting our bread on the waters.

I'll explain, but first let me say that my beliefs are born out of:

1. Nearly twenty years of active research interest in publishing, and especially in scholarly publishing

2. Thirteen years experience as founding coeditor and then editor emeritus of the Internet's first peer-reviewed electronic journal in the humanities, Postmodern Culture

3. Ten years experience as the first director of the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia

4. Eight years as member and then cochair of the MLA's Committee on Scholarly Editions

5. Three years as chair of the Text Encoding Initiative (TEI) Consortium, an interdisciplinary and international body of scholars and technical experts devoted to standards for the creation of machine-readable literary and linguistic texts.

I also speak to you as someone who got tenure in a top-ranked English department without authoring a book—tenure was based, instead, on article-length pieces, many of which were published electronically, and on applied research (in electronic scholarly publishing)—and as someone who was recently promoted to full professor and appointed Dean of the Graduate School of Library and Information Science at the University of Illinois, with a joint appointment as professor of English, still without a book.

The purpose of reciting this litany is not to impress you with my credentials—or to amaze you with my success in their absence—but rather to
demonstrate that the ideas I ask you to consider are based on direct, extensive, and personal experience, as well as research and experimentation.

To begin with, I do accept that there is a crisis in scholarly publishing in the humanities, and I agree with Professor Greenblatt (and the Association of Research Libraries, and others) that this crisis is to a significant extent the result of rapidly increasing prices for science, technical, and medical journals from commercial publishers. I also agree with Professor Greenblatt that the most straightforward solution to the problem this crisis poses for tenure and promotion is to accept several scholarly articles in place of a book. This solution requires relatively little adjustment: we are already familiar with the genre of the scholarly article, and we already value publication in this form; we would simply need to value it more. And from a business point of view, scholarly journals are more viable than Philip Lewis gives them credit for being, in his article titled “Is Monographic Tyranny the Problem?”2 Books sell once, journals sell three or four times a year—they are a renewable resource, economically speaking—and often journal profits carry book publishing in the presses that do both.

Another more or less traditional form of scholarly output that could be an alternative to the monograph is the scholarly edition. If the profession were of a mind to broaden its definition of tenurable work, broadening it in this direction would accomplish several things. First, it would reward the kind of work that is required, in each generation, to keep the cultural record up to date and in good repair. Second, it would promote this renovation at a time when great portions of that record are going to be transferred into electronic form, whether carefully selected and edited or not. A concerted effort to recognize and reward electronic scholarly editions might increase the odds in favor of the survival of the best, rather than the cheapest, texts.

Another proposal that has been discussed is subvention, and Professor Alonso has put forward the idea of MLA-sponsored subvention/prize committees, as a way of avoiding the appearance of departments buying publishing opportunities for their faculty.3 The problem with this idea is that even if every MLA member gave $10 a year to this cause, there would be funding for about 50 of these $7,000 subventions—and even if aggressive fundraising on the part of the MLA were to double those numbers, a hundred subventions a year would hardly make a dent in the situation, when university presses in the United States and Canada publish roughly 11,000 books a year.4 And you can bet that if the award is competitive, and there’s a chance of making a mistake by giving it to someone unknown, those 50 or 100 subventions will go to people whose reputations are already established, unless they are specifically restricted to untenured scholars.

Perhaps we simply can’t afford to publish in the way that we have been—in fact, perhaps the audience for humanities monographs is so small that this sort of book publishing can never be profitable. This assumption is at the core of Stevan Harnad’s 1994 “Subversive Proposal.”5 If you haven’t read this piece, and the responses to it, this would be a good time to review it. In summary, Harnad says that scholars are more interested in having their work read and used than they are in making money from it; that scholars have hitherto signed their work over to publishers because, in the print world, that was the only way to disseminate their work so that it would be read and used; that in the electronic world, authors don’t need to make this “Faustian bargain”; that when the audience for scholarship is small, there’s really no way that a publisher can afford to publish a book anyway; and that the reduced cost of electronic-only publishing for small audiences should be met in advance—by subsidies or through page-charges.

In a response to that article, in 1997, I said that all of this was right on the mark, but taking Postmodern Culture as an electronic-only example, nobody was offering to pay our costs up front, and we didn’t imagine it would be good for our submissions if we were to become the only humanities journal with page charges. So, often, there’s a sort of stalemate preventing the implementation of perfectly reasonable ideas about how to solve this crisis: one journal can’t initiate page charges if no others do; one department can’t change its tenure requirements if no others do; one faculty member can’t decide to skip the book...well, actually, one can, but it was frankly a very risky thing to do, and I wouldn’t recommend it. Still, perhaps these changes are more likely to come from the bottom up than from the top down, more likely to come from authors than from tenure committees, or journals, or publishers.

The Thematic Research Collection

If that’s true, then I predict that the genre of scholarship that will replace the book will be the thematic research collection. This genre has been independently identified by at least two different people before me—Daniel Pitti, in a talk given in Ireland in 1999 and Carole Palmer, who has a chapter on the subject in the forthcoming Blackwell’s Companion to Digital Humanities. The genre describes most of what we produce at IATH, and what other humanities researchers, often with less support, less funding, and less encouragement, are producing on University Web servers around the world. In a talk I gave at the University of Minnesota in 2001,7 I defined thematic research collections as:

1. Necessarily electronic (because of the cost of 2, 3, 8)
2. Constituted of heterogeneous data types (in other words, multimedia)
3. Extensive but thematically coherent
4. Structured but open-ended
5. Designed to support research
6. Authored (and usually multi-authored)
7. Interdisciplinary
8. Collections of digital primary resources (and they are themselves second-generation digital resources)

Thematic research collections offer the author all the benefits Professor Alonso ascribes to the book, in his recent “Editor’s Column” in PMLA, though sometimes in slightly different forms: “the choice of texts, the marshalling of sources and evidence, the construction of an argument that spans several chapters, the bibliographic research, the engagement with the readers’ reports, the reading of proofs, the choice of journals for review.” I could provide examples of IATH projects that have traced each of these steps—the Blake Archive, the Rosetti Archive, the Whitman Archive, the Valley of the Shadow, and others. And I’m sure the faculty who have assembled, edited, annotated, and analyzed these thematic research collections would agree that they are the result of what Professor Alonso, in the case of the scholarly book, describes as a “protracted and somewhat enigmatic process to which many people contribute, sometimes unknown to them.”

Now, I would not for a minute suggest that these thematic research collections are less expensive to produce than a scholarly book—far from it. Nor would I suggest that they are an easier nut to crack, from the point of view of the business of publishing. But in spite of both of those things, I think they may be more viable, because they have something that most scholarly books do not, namely an audience. It’s hard to sell five hundred copies of most humanities monographs; few sell in the thousands. And yet, these Web-based projects, on relatively esoteric subjects, receive thousands of visitors each day, serve up gigabytes of their content to avid users each week, and reach readers of all ages, inside and outside academia, and around the world. The only problem is that they’re free.

**Enlarging the Audience for Humanities Scholarship**

Or maybe that’s not such a problem. Let’s take the case of Postmodern Culture again. It has always been a free electronic journal, but since the mid 1990s, it has also been a licensed electronic journal—you can get it for free, or you can pay for it. Most individuals get it for free. Most paying customers are institutions, who buy it as part of Project Muse. You could argue that they’re just buying that package, not the title, and that’s perfectly true—but Postmodern Culture is one of Muse’s top 10 most heavily used titles, so clearly the end users—who could be using the journal on the free site—are choosing to use it as part of the licensed resource. Why? Better searching, for one thing, and searching in the context of a hundred other humanities journals.

More to the point, perhaps, given the stalemate problem, why did Johns Hopkins agree to take us on, and allow us to continue distributing the journal for free at the same time they were licensing the journal? The answer is simple: we came to them with an established audience, and we had established that audience by being free. Another reason Johns Hopkins took us on was that Michael Jensen, one of the few visionaries in scholarly publishing these days, was at that time the manager of Project Muse. When we signed with Muse, we had 2,500 subscribers to the e-mail list through which we announced new issues and we were getting a quarter of a million visits a year on the Web. By anybody’s standards, that’s a large audience for a scholarly journal, and especially for one that published articles with titles like “Flogging a Dead Language: Identity Politics, Sex, and the Freak Reader in Acker’s Don Quixote.” In fact, I think that a solution to this crisis is, plain and simple, to reach a larger audience. We tend to condescend to the general reader, and we court her out when it comes to our mental construct of the audience for humanities scholarship—and yet, believe it or not, this is an actual e-mail I received one day in the mid-1990s, from a reader of Postmodern Culture:

Dear Mr. Unsworth: I’m a union teamster living in rural Vermont so I don’t have a lot of access to the sort of stuff you have in your journal and you provide access to from your Web site. Our local library is swell, computerized too, but a computer search under postmodernism or poststructuralism or Derrida or Baudrillard or Jameson produces zero hits. Thank you.

Would this rusticated teamster buy up the latest book by Jameson? Maybe, if he could get his hands on it in Rutland, or wherever he is, and maybe not—but he is clearly interested in the subject, and looking for the content. Maybe we could enlarge the audience for humanities scholarship, not by dumbing it down, but by making it more readily available. Maybe if we did that, scholars would find an audience first, and a publisher second, instead of the other way around. And maybe in that world, the risk to publishers would be less, because the demand would already be demonstrated. Could we
peer review in this world? Of course—and it might then be perfectly clear why we should conduct peer review independent of a decision to publish. Could we give away and charge for the same thing, in different venues? Yes, if the benefits to paying customers were real. Could books still exist? Yes, but they might often be byproducts of other activity, for example summaries or extracts of research conducted in the course of building thematic research collections. Could we collect and publish the collections themselves? Absolutely—the only clear business lesson that we have learned from electronic scholarly publishing to date is that size matters, size sells, and size is achieved by aggregation. Collections of collections, rationally organized and critically selected, would make perfect sense, and their individual components, freely available on the Web, wouldn’t cannibalize the market, because they wouldn’t have the same scale or reach.

Enough. I know that these ideas fly in the face of what we all know about the business of scholarly publishing, about the audience for humanities research, about the forms that research naturally takes, and so on. But the simple truth is that the crisis we’re discussing is the lack of an audience, and I know that the audience exists—we just need new genres, new business models, and the courage of our convictions as scholars and publishers, to reach them.

—Copyright 2003 John M. Unsworth

1 Stephen Greenblatt, then-President of the MLA, wrote a letter to MLA members dated May 28, 2002. Greenblatt encouraged members to discuss within their departments the scholarly communications issues affecting the ability of junior faculty to meet rising expectations for tenure. The letter was reprinted along with responses from four scholars as “Open Forum: Scholarly Publishing and the Tenure Process,” Literary Imagination 5, no. 1 (Winter 2003): 151–164.


PubMed Central—Three Years Old and Growing Stronger
by Edwin Sequeira, National Center for Biotechnology Information, National Library of Medicine
<sequeira@ncbi.nlm.nih.gov>

PubMed Central (PMC) is the National Library of Medicine’s (NLM) digital archive of medical and life sciences journal articles. It was conceived in the spring of 1999 when Harold Varmus, then director of the National Institutes of Health (NIH) of which NLM is a part, proposed that NIH create and manage an open archive of research papers in the life sciences. Many of the early exchanges about the proposal within the publishing community made it sound as if revolution was in the air. The reality, however, is that PubMed Central represents evolution not revolution. PMC is here to stay, but it does not spell disaster for academic societies and other publishers.

It is probably no coincidence that a number of traditional subscription-based journals began offering free online access to their back issues within months of NIH’s open-archive proposal. Happily, what may have begun as a preemptive move against the NIH initiative now seems to be a trend that more and more journals are following. Most people who know about PMC probably associate it only with free access to journal articles, but PMC is equally about archiving and the integration of journal literature with other research resources, such as genetic sequence data.

NLM is the world’s largest biomedical library, with a collection in excess of 6 million items. As a national library, NLM is committed to preservation and to public access. PMC extends to digital journal literature what NLM has done for more than a century with the print literature. NLM’s National Center for Biotechnology Information (NCBI), which designs and operates PMC, also operates PubMed/MEDLINE (NLM’s 12 million-plus citation database that dates back to the mid-1960s), GenBank (the genetic sequence database), an online Books database, and various other data resources for biomedical professionals. Access to most of these resources is provided through an NCBI service known as Entrez, which makes seamless links between related objects across the databases.

PubMed Central’s Basic Policies
A 15-person PubMed Central National Advisory Committee <http://pubmedcentral.gov/about/nac.html> advises the Director, NIH; the Director, NLM; and the Director, NCBI, on the content and operation of PMC. Committee members are appointed by the Director, NIH, and include scientists, librarians, and publisher representatives.

Eligibility to Participate in PubMed Central
To participate in PMC, a journal must either be covered by a major indexing service, such as MEDLINE, Agricola, Biosis, or EMBASE, or have three members on its editorial board who currently are principal investigators on research projects funded by major nonprofit agencies, such as NIH, in the U.S. or abroad. These criteria provide a simple way to ensure that a journal has a reasonable degree of editorial oversight.

In addition, a journal must be able to supply its current content to PMC in the form of fully tagged SGML or XML for the text of articles and supply original high-resolution image files for graphics. The journal is also encouraged to deposit a PDF (print format) file to complement the SGML/XML version of an article, and supplementary or supporting research data (video clips, data sets, etc.) where available.

Access Policy
PMC’s open access principle is that a journal may restrict access to its content for some publisher-defined time after publication, beyond which period it is freely available to everyone forever. This open-access delay is generally a year or less. There is a class of articles in PMC, labeled “PubMed” articles, for which full text is viewable only at the journal’s own site, even though the journal has deposited the full text in PMC. Journals with PubMed articles make the same commitment to providing permanent open access to all content they have deposited in PMC (with the usual provision for initial access restrictions). Over time, PubMed articles can also become viewable directly in PMC.

Where Is PubMed Central Today?
The PMC Web site <http://www.pubmedcentral.gov/> made its debut in February 2000 with one issue each of Proceedings of the National Academy of Sciences of the USA (PNAS) and Molecular Biology of the Cell. As of May 2003, PMC contained almost 100,000 articles from over 120 journals. In the same month, users retrieved the full text of articles 700,000 times in PMC and were “PubMedbed” to a journal site for full text another 70,000 times.
Searching and Cross-Database Linking

Users of the PubMed/MEDLINE bibliographic database, which includes abstracts of PMC articles, can link directly from a PubMed abstract to the corresponding full text in PMC. The full text of all PMC articles (including PubLink articles, which can only be viewed at the journal site) is also searchable in the Entrez system. The standard PMC search technique is labeled “SmartSearch” reflecting the fact that it is based on an automated analysis of the title, abstract, and full text of each article. SmartSearch is intended to increase the relevance of one’s search results. It includes intelligent phrase recognition and does not search every word in an article as a simple full-text search would do (although it is also possible to do the latter if one wishes).

PMC also uses automated analysis to identify terms in an article that most likely are organism names. These names are then used to provide bidirectional links between the article and related records in the Entrez Taxonomy database. Whenever possible, a search using a common name for an organism is translated to its scientific name, e.g., “yeast” —> “Saccharomyces cerevisiae.” Similar automated detection and linking applies to accession numbers in the text for molecular sequences and structures registered in other Entrez databases. Advanced search features let you search just selected parts of articles, such as the Methods section, figure and table captions, or section headings.

The Links pop-up menu associated with each article on a search result page (Figure 1) gives you immediate access to the above-mentioned links as well as others, such as a “cited in” list of all articles in PMC that have cited the current article, or a “Referenced articles” link that produces a list of PubMed citations and abstracts of articles in the References section of the PMC article. You can also display a combined set of links for a group of articles by using the checkbox alongside each article to select the articles and then using the “Display” option at the top of the result page. The References section of a PMC full-text page also has direct links to the full text of cited articles that are in PMC, as well as to the full text of many other journals that participate in NLM’s LinkOut program.

All of the links in PMC and the other Entrez databases are generated automatically and dynamically, so that when new related content is added to PMC or another database, the corresponding links appear either immediately or overnight in the worst case. This integration of PMC with the wide array of bibliographic and factual databases in Entrez means that a user can often link directly from an article to updated or more detailed data, even for rapidly developing subjects.

Building and Managing a Digital Journal Archive

In the three years that PMC has existed, NLM has gained a wealth of experience in the practical realities of building and managing a digital journal archive. Figure 2 is a simplified view of what PMC does with data submitted to the archive. It can be summarized in four words: convert, load, render, validate. Data validation is a part of each of the other activities, as explained in a subsequent section of this article.
Convert
The high-resolution image files supplied by a journal, which are generally quite large, are converted to a format more suitable for faster downloading and display on the Web. At the same time, the SGML/XML files are converted to a common XML format or DTD (Document Type Definition). This conversion retains both the content and definition of the different parts of an article from the source files, but may change the way individual elements are tagged. For example, if one journal tags an author name as author and another as auth and the third as au, they may all get translated to author-name in PMC.

There are a large number of DTDs in use for creating SGML/XML files of the full text of journal articles. PMC alone receives content in almost a dozen different DTDs. Something the PMC team learned early is that it is not practical to work with multiple DTDs in an active archive. Because NLM cannot force participating journals to all use the same DTD, it uses conversion to achieve a similar effect. Converting to a common DTD provides tremendous efficiencies. The complexities of handling diverse DTDs are dealt with once, when a file enters the archive. From there on, the uniform data format means simpler software for managing and retrieving data from the archive, as well as easier redistribution of the data to other archives.

Load
All the journal’s source files are loaded into the PMC data repository, together with the converted PMC XML files and Web display images. Part of this repository forms the permanent PMC archive, and part of it is the publicly accessible PMC database, as shown in Figure 2.

Render
PMC’s online display pages are created dynamically from the objects in the PMC database. For instance, a PMC table of contents (TOC) page for an issue is not a static HTML page that is saved somewhere in PMC. Each time a user requests a TOC page, it is created by collecting all articles in the database that were tagged in the source SGML/XML files as belonging to that issue. In the case of a journal that releases articles individually as they are accepted, the issue TOC is automatically updated as soon as an article is loaded into the database. Other PMC pages such as “cited in” lists are also automatically updated in the same way. Full-text article displays are also created dynamically, directly from the PMC XML. PMC’s software has been tuned so that the response time for dynamic rendering of articles is not noticeably longer than for displaying a static HTML page.

Data Validation
The files deposited by a journal are checked for accuracy at all three of the above stages. A number of automated tests are done at the conversion stage for what may be termed mechanical accuracy: that the syntax of the SGML/XML conforms to its corresponding DTD, that an image file exists for each graphic referred to in the SGML/XML, and so on. Additional automated checks are done during the loading stage, primarily for consistency of journal- and issue-level data across
all the articles in a batch. Then, after an issue is loaded into the database, a PMC reviewer spot checks the online issue in PMC similar to what a copy editor would do, looking for discrepancies between the online version and the print copy or other authoritative source.

PMC's data validation, which almost always follows a journal's own production quality checks, adds to the overall level of archival file quality. A journal is asked to correct all errors found in PMC's automated checks because the files cannot be processed with them. Correction of "copy errors" is limited to the more serious cases. For practical reasons, we do not ask a journal to correct minor formatting or typographical errors that do not affect the readability and understanding of the content, unless the journal chooses to correct them.

**PubMed Central's Archiving Philosophy**

PMC's archiving philosophy, stated simply, is:

- **Store all material in the archive in formats that are portable, self-defining, and minimally dependent on specific software or equipment.**
- **Use the material continuously to confirm that, in fact, it continues to be usable. (Otherwise, why preserve it?)**

The first point underlies the choice of SGML and XML as archival data formats, rather than HTML, the lingua franca of the Web. Although someone looking at a page of XML for the first time may find it bewildering, XML (or SGML) is actually relatively easy to decipher. In mechanical terms, an XML file is a plain text file. It can be read and interpreted by a human. To do so, one only needs a reasonably well-documented DTD, which serves as the codebook for the XML, so to speak. It is a safe bet that wherever technology goes in the next 50 or 100 or more years, an XML file will still be interpretable and usable.

XML also preserves the structure of an article by explicitly identifying individual elements, such as the article sections and section heads, or the journal, volume, issue and other details for each bibliographic reference. This explicit definition supports both accurate reproduction and innovative reuse of content. For instance, it provides a precise context for the words in an article: words appearing in a section heading versus in a figure caption; an author of the main article versus an author of a referenced article. That information can be used to build a search system that gives added weight to terms that appear in figure captions or section headings of articles, or allows the identification of reagents listed only in the methods section of an article. HTML, on the other hand, is not much more than a blob of text with embedded links. Variations in font face and style are useful for presentation but do not provide a reliable defined structure. By virtue of being geared to presentation, HTML also includes a lot of data extraneous to the actual article content. Furthermore, the graphics on an HTML page exist as separate files that must be saved with it, if an article is to be recreated in its entirety, and image files that have been compressed for online transmission usually are of lesser quality than the original images.

In line with the notion of self-defining files, PMC requires that certain data elements are present, consistently, in the SGML/XML file for every article in PMC, even if the corresponding DTD does not require them. These elements include the journal ISSN and a standard journal title; volume, issue, pagination, and publication date; publisher; and a copyright statement, where applicable. Surprisingly, some commonly used DTDs do not provide for explicit tagging of some of these elements. In other cases, a journal may not care that its files lack a journal issue date or a copyright statement, in the belief that it is sufficient if that information appears on its Web site.

For proprietary file formats such as PDFs and Microsoft Excel spreadsheets, PMC's current approach is to assume that other technology groups will continue to provide tools that make them usable because of their ubiquity.

**Copies of the Archive**

NLM believes that the best way to ensure the durability of an electronic archive is to use it constantly. Because PMC creates its online displays directly from the archival files, as illustrated in Figure 2, readers are confirming and validating the quality of the archival files every time they retrieve an article. This adds another level of assurance to NLM's own testing and quality control procedures.
The Journal Archiving and Interchange DTD Suite
As mentioned in the section Building and Managing a Digital Journal Archive, all full text is stored in PMC in a common XML DTD format. This DTD, known as the pmc-1 DTD, was first developed in the latter part of 2000. It began as a DTD that could accommodate content from the two publisher DTDs in which PMC was receiving files at the time and was updated and extended as new DTDs made their way into PMC. By late 2001 it was clear that frequent updating to accommodate a new DTD was not viable for the long term. With the help of an XML consulting group, NLM began developing a new DTD that would make allowance for a much broader range of incoming DTDs from the start. Near the end of this effort, NLM was approached by the Andrew W. Mellon Foundation and the Harvard University Library, which—under Mellon’s e-journal archiving program—had been studying a variety of DTDs and the feasibility of having a single archival DTD for journal literature.

Harvard and Mellon proposed extending the new PMC DTD so that it could be used for archiving journals in all disciplines, rather than just the life sciences that NLM had focused on up to that point. The proposal seemed to be both reasonable and feasible. Following through on the idea, the developers of the PMC DTD, aided by a second XML consultant hired by Harvard/Mellon, analyzed all the major DTDs in use for journal literature to ensure that the expanded PMC DTD would be comprehensive and flexible enough to serve as a universal archival container. The result is the Archiving and Interchange DTD Suite, which takes a modular approach to building DTDs. The suite is a set of building blocks from which any number of specific DTDs can be created with relatively little effort.

Using the suite, NLM has created the Archiving and Interchange DTD <http://dtd.nlm.nih.gov/>, which will replace the pmc-1 DTD as the foundation of the PMC archive later this year. This Archiving DTD is of necessity quite flexible so as to accommodate input from a large number of other DTDs. NLM has also released a much more restrictive Journal Publishing DTD which can be used by a journal to mark up its content in XML. Both DTDs and the underlying suite are in the public domain. The Archiving and Interchange DTD can serve not just for archival purposes but also as a standard format for moving content between publishers, content aggregators (redistributors), and archives. The modular nature of the suite also allows it to be used to build DTDs for content beyond journals, by adding the necessary modules to cater to the specific characteristics of, say, books.

Reaction to the new DTDs has been very positive. Already, one journal production vendor that supplies PMC data on behalf of several journals is using the Journal Publishing DTD for its newest contributor. At least one large online journal site and another archiving organization are studying the feasibility of adopting the DTDs, and several small publishing operations have begun working with the Publishing DTD. NLM is in the process of establishing a panel of outside XML experts to advise NLM as changes and updates are called for to keep the DTD relevant to publishing and archiving operations.

Back-Issue Scanning Project
A year ago NLM began planning a project to scan, cover to cover, the complete run of back issues of a select set of journals. NLM offered to scan any issues of a PMC journal that are not already available in electronic form, in return for permanent rights to archive and distribute the scanned material freely. Almost all the current PMC journals that have pre-electronic issues are participating in the project, as are the 20-plus specialist journals of the BMJ Publishing Group, whose current content will be added to PMC later. Each journal will receive a complete electronic copy of its material, at no cost, to do with as it chooses.

NLM will archive the original scanned page images (TIFF files) and will make PDF files of all articles and other scanned content freely available through PMC. Grayscale and color graphics in the articles will be reproduced in the PDF files as true representations of the original pages. OCR text will be generated from the scanned pages to support full-text searching. For articles not already in PubMed, abstracts will be keyed as XML-tagged records for inclusion in both PMC and PubMed.

Production scanning began in May 2003 and is scheduled to last a little over a year for the first phase, which will cover an estimated 3.5 million pages. The first scanned issues should be available in PMC sometime this summer. For the near term, NLM will offer this service to other journals that join PMC.

Help NLM Expand PubMed Central
NLM is committed to expanding PMC and adding more features to meet the needs of the scientific community. As has been true for other NLM services, including PubMed and LinkOut, academic libraries can contribute to the continuing development of PMC. NLM invites your ideas and support and hopes that you and your colleagues will persuade more journals to join PMC. You can visit the PMC site for an overview of what is involved from a publisher’s perspective <http://www.pubmedcentral.gov/about/pubinfo.html> or send your suggestions and questions to PMC <pubmedcentral@nih.gov>.
ARL SALARY SURVEY HIGHLIGHTS
by Mark Young, Research Assistant for Statistics and Measurement

The recently published ARL Annual Salary Survey 2002–03 reports that ARL librarians’ salaries experienced one of the smallest increases in the survey’s history. The combined median salary for U.S. and Canadian ARL librarians rose at a rate of 1.8% over the past year, a lower increase than in any ARL Salary Survey since 1980. This small increase out-performed inflation in the U.S., however, where the Consumer Price Index (CPI) rose 1.5%, but not in Canada, where the CPI increased 2.1%. The 2002–03 survey reports on 9,469 professional staff members for the 114 ARL university libraries (including law and medical libraries), and 3,804 staff members for the 10 nonuniversity ARL libraries. Overall, the median university library salary was reported at $51,636; the median nonuniversity library salary was $65,289.

Median beginning salary increases also slowed, but not as sharply as overall salary increases. The median beginning salary in ARL university libraries rose by $1,000 to $35,000, a 2.9% increase after four straight years of rising by 3% or more. ARL nonuniversity libraries saw their median beginning salary rise only 1% (to $34,739), a strong contrast to the 8% increase this group experienced a year ago.

The purchasing power for the Canadian dollar continues to weaken. Canadian university libraries for FY 2002–03 recorded a median salary of $42,657, almost one-fifth less than the U.S. median of $52,789.1

The Pacific, New England, and Middle Atlantic regions again had the highest average salaries in the United States. The gap between salaries in private U.S. ARL university libraries and those paid in publicly supported U.S. university libraries was 6.4%, the same as in 2001–02. Libraries with staff levels higher than 110 have the highest average salary, $58,828, compared to $58,139 for libraries with between 75 and 110 staff. Libraries with staff of 50–74 professionals paid an average salary of $53,192 and those with staff between 22 and 49 paid $54,579. The difference in salaries between the highest paying cohort and the lowest paying cohort is $4,249, about 24% smaller than last year’s gap of $5,583.

The gender gap in ARL university salaries still exists, and may even have widened over the past year. The average salary for men at all ARL institutions was $58,295, while women’s salaries averaged $53,953—94.1% of the men’s average salary; this gap is virtually identical to the gap in 2001–02. Average salaries for men surpass those of women in 18 of the 27 job categories that ARL tracks, compared to 15 in 2001–02. Differentials in experience do not explain this phenomenon; there are several categories in which women average more experience but lower salaries, including Director of Libraries. This pattern is also repeated for minority librarians: the average salary for minority men is higher than that for minority women in nine of the ten experience cohorts. As in 2001–02, the average salary for male university library directors (55 men out of 111 directorships reported) surpassed that of female directors by 2.7%. Overall, ARL university libraries are approximately 64% female, a figure that has remained relatively consistent since 1980–81.

Minority librarians comprise 12.4% of the professional staff in U.S. ARL university libraries, with the number of minorities in managerial or administrative positions being lower (5.2% are directors; 6% are associate or assistant directors; and 10.3% are branch librarians). Women comprise 69.3% of minority staff.

The ARL Annual Salary Survey 2002–03 is available for $60 to member libraries and $120 to nonmembers (plus shipping and handling), and is available on standing order. To order online, visit <http://www.arl.org/pubs/cat/order/>. For more information, contact ARL Publications at <pubs@arl.org>.

1 Canadian salaries were converted to U.S. dollars at the rate of $1.56678 Canadian to $1 U.S., the monthly noon exchange rate published in the Bank of Canada Review for the period July 2001–June 2002.

<table>
<thead>
<tr>
<th>ARL Academic Librarians, FY 2002–03*</th>
<th>Men</th>
<th>Women</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Filled Positions</td>
<td>2,832</td>
<td>4,959</td>
<td>7,791</td>
</tr>
<tr>
<td>Average Salary</td>
<td>$58,295</td>
<td>$54,717</td>
<td>$56,018</td>
</tr>
<tr>
<td>Average Yrs Exp**</td>
<td>16.8</td>
<td>16.8</td>
<td>16.8</td>
</tr>
<tr>
<td>Total Number of Minorities (U.S. only)</td>
<td>265</td>
<td>599</td>
<td>864</td>
</tr>
<tr>
<td>Minority Average Salary (U.S. only)</td>
<td>$54,469</td>
<td>$52,035</td>
<td>$52,782</td>
</tr>
<tr>
<td>Minority Average Yrs Exp (U.S. only)**</td>
<td>14.1</td>
<td>15.0</td>
<td>14.7</td>
</tr>
<tr>
<td>Total Number of Directors</td>
<td>55</td>
<td>56</td>
<td>111</td>
</tr>
<tr>
<td>Average Salary of Directors</td>
<td>$149,041</td>
<td>$145,067</td>
<td>$147,036</td>
</tr>
<tr>
<td>Average Yrs Exp of Directors (filled positions)**</td>
<td>29.4</td>
<td>29.9</td>
<td>29.6</td>
</tr>
</tbody>
</table>
ARL ACTIVITIES
Kaylyn Hipps, Editorial and Research Associate

ARL NAMES NEW PROGRAM OFFICER FOR TRAINING AND DIVERSITY

Jerome Offord, Jr., was appointed ARL Program Officer for Training and Diversity effective April 7, 2003. He manages the ARL Diversity Program, which houses several national recruitment and professional development programs: the Initiative to Recruit a Diverse Workforce, the Leadership and Career Development Program, and the Career Resources Service. In addition, Jerome serves as an adjunct faculty member for the Office of Leadership and Management Services (OLMS). In this capacity, he delivers workshops, training sessions, and seminars to the library community.

Prior to working at ARL, Jerome was Director of Finance and Operations for Us Helping Us, Inc., a small, Washington, D.C.-based nonprofit organization. He has experience as a student affairs professional at George Washington University and American University. Jerome has been heavily involved in diversity programming and serving diverse communities throughout his career and is quite familiar with the culture of academic institutions. He also has significant grant-writing experience. Jerome holds a B.S. from Lincoln University in Missouri and an M.S. from Colorado State University. He began the M.L.S. program at the Catholic University of America this summer. Jerome may be reached at the ARL office or by e-mail <jerome@arl.org>.

TRANSITIONS

California, Los Angeles: Gary Strong, Director of Queens Borough Public Library, was named University Librarian effective September 1, 2003.

Library of Congress: Deanna Marcum was named Associate Librarian for Library Services, effective August 11. She is currently President of the Council on Library and Information Resources (CLIR).

North Carolina: Joe Hewitt announced his intention to retire as University Librarian effective June 30, 2004.

Pennsylvania: Paul Mosher resigned his position as Vice Provost and Director of Libraries. Carton Rogers, Associate Director of the Library and Director of Information Processing, is serving as Interim Director.

Texas A&M: C. Colleen Cook will serve as Dean of the Texas A&M University Libraries on an interim basis when Fred M. Heath relinquishes the post on July 31, 2003. She is currently Associate Dean of Libraries and a professor of library science at Texas A&M.

Texas at Austin: Fred M. Heath was named Vice Provost and Director of General Libraries effective August 1, 2003. He has served as Dean of Libraries at Texas A&M University since 1993.

Waterloo: Mark Haslett was appointed University Librarian effective May 1, 2003. He served as Associate Librarian, Information Services and Systems, since 1996.

Other Transitions

American Council of Learned Societies: Pauline R. Yu was named President effective summer 2003. She was previously Dean of Humanities in the College of Letters and Science and Professor of East Asian Languages and Cultures at the University of California, Los Angeles.

Committee on Institutional Cooperation: Susan Singleton was named Director of the CIC Center for Library Initiatives. She was previously Executive Director of the Missouri Library Network Corporation.

Government Printing Office: Judith Russell was named Superintendent of Documents and George A. Taylor was appointed Deputy Public Printer effective January 6, 2003.

JSTOR: Mike Spinella, formerly of the American Association for the Advancement of Science, was named Executive Director effective February 24, 2003. He replaced Kevin Guthrie, who now chairs the JSTOR board of trustees.

Library of Congress: Hwa-Wei Lee, former Director of Libraries at Ohio University, was named Chief of the Asian Division.

HONORS

Ross Atkinson, Cornell University, received the 2003 Association of College and Research Libraries (ACRL) Academic/Research Librarian of the Year award.

Heather Joseph, President and COO of BioOne, received the 2002 Association of Learned and Professional Society Publishers (ALPSP) Award for Service to Not-for-Profit Publishing.

Werdy Pradt Lougee, University of Minnesota, received the 2003 Association of College and Research Libraries (ACRL) Hugh C. Atkinson Memorial Award.

North Carolina State University Libraries received the 2003 American Library Association/Information Today "Library of the Future" award for a Web-based library orientation program that helps students learn to use library services quickly and effectively.

Catherine Quinlan, University Librarian at the University of British Columbia, received the 2002 Outstanding Alumni Award from the School of Library and Information Studies, Dalhousie University.
ARL Calendar 2003

July 24–26  Sound Savings: Preserving Audio Collections
Arlington, Virginia

July 28–29  ARL Board Meeting
Washington, D.C.

August 13–15  Advanced XML: Data Transformation with XSLT
Charlottesville, Virginia

September 8–9  Exposing Hidden Collections
Washington, D.C.

September 22–25  Library Leadership for New Managers Program
Washington, D.C.

October 3–4  New Ways of Listening to Library Users: Tools for Measuring Service Quality
Washington, D.C.

October 7–8  Leading Change
Washington, D.C.

October 14–16  ARL Board and Membership Meeting
Washington, D.C.

October 17  Scholarly Tribes and Tribulations: How Traditional Practices in the Disciplines are Driving Technology in Different Ways
Washington, D.C.

November 4–6  Library Management Skills Institute I: The Manager
Los Angeles, California

December 8–9  CNI Fall Task Force Meeting
Portland, Oregon

Online Lyceum
Can’t make it to our in-person events?
Take a look at our Online Lyceum Web-based course offerings at <http://www.arl.org/training/lyceum.html>.