Evidence Based Library and Information Practice

Commentary

Digital Archiving of Primary Research Data

A. James Bothmer, MALS, AHIP
Assistant Professor and Director
Health Sciences Library/Learning Resources Center
Creighton University
2500 California Plaza
Omaha, NE 68178
E-mail: jbothmer@creighton.edu

Robert P. Heaney, MD
John A. Creighton University Professor and Professor of Medicine
Creighton University School of Medicine
2500 California Plaza
Omaha, NE 68178
E-mail: rheaney@creighton.edu

Ramon M. Fusaro, MD, PhD
Adjunct Professor of Preventive Medicine and Public Health
School of Medicine and Adjunct Professor of Pharmaceutic Sciences
School of Pharmacy and Health Professions
Creighton University
2500 California Plaza
Omaha, NE 68178
and
Professor of Internal Medicine (Dermatology)
College of Medicine
University of Nebraska Medical Center
984360 Nebraska Medical Center
Omaha, NE 68198-4360
E-mail: rfusaro@unmc.edu

Received: 09 July 2007 **Accepted:** 10 July 2007

Only a relatively small portion of the data generated from research projects performed at universities ever sees the light of day. Final successful experiments are usually the only data that are published. For economic reasons, journals publish minimal details of a study's methodology needed for others to replicate the findings. Research is a cultural and scientific treasure of intellectual effort; somehow, therefore, universities should preserve all raw data from research conducted within their domains, successful, unsuccessful, published, and unpublished. Academic health sciences libraries need to take the lead in this effort by creating and linking Institutional Repositories (IR) to form a true national database of all intellectual experimentation. These efforts can be implemented without changing the status quo of traditional publishing and they could be introduced by reallocating existing resources without extensive technical development. These publications would still need internal universities' faculty peer review to assure the integrity and quality of intellectual accomplishments.

Digital technology has affected how scholars disseminate and preserve their research. It is now economically possible to archive information with digital technology making it possible for libraries to catalogue it for greater access by the scholarly community (Lynch). The current technology also makes it possible to include appropriate restrictions for prepublication concealment, patentability and patient confidentiality. Institutional Repositories were devised as a solution for preserving and making accessible the scholarly output of an institution's researchers (Lynch; Harnad; Crow). Institutional Repositories are defined here "as a formally organized, digital collection of the intellectual output of an academic campus community and comes directly from the faculty, staff, and students of the university" (Singarella 2). There are essentially two schools of thought regarding an IR. One articulated by Lynch argues that an IR serves to disseminate 'grey literature' including such documents as pamphlets, bulletins, visual conference presentations, and other materials that are typically not included in traditional publications (Lynch). The alternate view expressed as early as 1995 indicated that an IR could effectively take the place of traditional publishing, or at least be a competitive entity (Harnad).

The purpose of this opinion piece is to argue that a national network of IRs could be used to archive and preserve published and unpublished raw data in institutional resources. Access to raw data could prevent unnecessary replications of investigations, provide data for university internal review of faculties' research, and permit access by other scientists to details of all research data that could be highly useful in light of newer insights and concepts. The authors believe it is a regular occurrence for researchers to wish they had access to unpublished raw data in a given publication, either to analyze it in a different way or to utilize measurements made in a given study but not published because they did not seem to bear on the primary study's outcomes. The availability of raw research data of reported clinical studies would increase the plausibility for either acceptance or rejection of previous proposed findings in evidence based medicine studies.

In order to make a case for harvesting research data (published or unpublished), we offer two examples that illustrate this need: (a) the need for access to the actual measured values if one is to reanalyze the

experiments of other investigators using different models and/or testing different hypotheses; and (b) the need for publications to contain full methodological details, the disclosure of which would help explain or contradict published findings. Without such access by other investigators, erroneous conclusions may be enshrined in the literature effectively forever. In the first example, below, the error in an earlier publication would likely ultimately have been found out, but only at the cost of unnecessary duplication of work already done, but misinterpreted. In the second example, a promising treatment might have been lost forever because of the inadvertent use, in this instance, of deteriorated chemical reagents, unrecognized as such by the investigators.

A pertinent archival retrieval experience of the first example comes from a recent in print publication describing the relationship of serum vitamin D concentration and that of its principal metabolite 25-hydroxy-vitamin D [25(OH)D] (Hollis et al., "Circulating"). The authors had used a type of curve fitting that led them to conclude that the concentration of 25(OH)D would reach a maximum at a certain serum vitamin D level and rise no further. One of us [RPH] had generated a similar set of measurements and wished to subject both sets to a different kind of curve fitting. The raw data were obtained by personal contact with the original investigator, and when combined, the two datasets constituted essentially the totality of the world experience and hence had unique value for unraveling important questions in a rapidly expanding field of investigation. As it turned out, the data actually showed that serum 25(OH)D rose without limit as serum vitamin D rose, a finding at variance with the original authors' analysis. Even more significant, the combined set allowed further and crucial insights into vitamin D metabolism that had not been apparent within the individual datasets (Hollis et al., "25-hydroxylation").

The second example is a complex investigation lasting over two decades. The investigation concerns the use of sequential bedtime skin applications of two separately stored chemicals (dihydroxyacetone {DHA} followed by lawsone) that produce a melanoidins-sunscreen in the top keratin layer of the skin that lasted for 1-2 weeks. (Fusaro). This skin-bound sunscreen gives sunlight protection for both normal persons and photosensitive persons who are either allergic to or intolerant of ultraviolet (UV-A or UV-B) wave lengths of sunlight without loss from perspiration. A clinical trial (Rice) of the above medication procedure reported complete sunlight protection of thirty photosensitive patients without any failures of protection. One of us (RMF) contacted the author of this study in order to obtain access to unpublished data. These data showed that the majority of the patients tested the limits of their sunlight tolerance and were also protected against UV-B sunburn for 6-8 hours a day (SPF 18-24) over a 7-month period (Fusaro and Rice, "Maillard Reaction"). In the study by Rice (Rice), the author tested the use of previously stored mixtures of DHA/lawsone. The same photosensitive patients received no significant sunlight protection as the compounds had degenerated and use of the mixture in patients was abandoned. A previously classified U.S. Army study (Fitzpatrick and Pathak) reported using two different methods for skin applications of stored mixtures of the same reagents and different sunburn testing procedures in two groups of the four volunteers that failed to provide significant sunburn protection compared to a PABA sunscreen (Fitzpatrick and Pathak). However, the authors later reported (Pathak, Fitzpatrick and Frenk) the four volunteers as one group in spite of the differences in topical applications of the DHA/lawsone mixture and two different

sunburn testing procedure. Two U.S. Senators were successful in obtaining the raw data from this classified study (Fitzpatrick and Pathak). The variable inadequate protective results of this U.S. Army clinical trial of four individuals gave concern that the DHA/lawsone mixture had chemically deteriorated while in storage. This deterioration was noted over a decade later by one of the U.S. Army's original investigators for the first time in a single paragraph about the use of DHA/lawsone mixture as part of a complete review of all available sunscreens (Pathak). The review's author noted that the deficiency of the stored mixture of the two compounds in their earlier study had changed color prior to its skin application and was washed off of the test subjects when they later bathed in water (Fitzpatrick and Pathak). This confirmed the chemical deterioration of the mixture and the failure to produce a skin-bound, melanoidins-sunscreen with high SPF sunburn protection in the U.S. Army study.

From personal contacts with the vitamin D investigators involved in the first example and a relentless in-depth search over two decades of published and unpublished data of the long-lasting sunscreen in the second, the two archival investigations uncovered the raw data needed to produce new conclusions.

Huge barriers exist to the creation of a nationally linked network of institutional repositories. For example, the Centers for Disease Control and Prevention (CDC) have created a large number of datasets from research. Until recently these huge and valuable datasets have not been easily identifiable. Metadata was developed for these CDC datasets creating a system that allows researchers to search over 95% of the databases within CDC. Unfortunately, the metadata and system is available only to CDC researchers thus marginalizing its usefulness (Matters et al). Another barrier to populating an IR with unpublished research data includes resistance and reluctance of some researchers to contribute their raw data. Davis has written that each discipline has a normative culture, largely defined by their reward system and traditions. If the goal of an IR is to capture and preserve the scholarship of one's faculty, this institutional cultural diversity will need to be addressed.

One strategy to overcome a reluctance to deposit research in an IR is to implement embargoed access restrictions (i.e., prepublication concealment, etc.), which will help ensure acceptance and compliance by research investigators, as it will protect their creativity. Overcoming these barriers will take time as well as open dialogue and collaboration among researchers, librarians and IT staff.

After reviewing issues of journals such as the *Journal of Clinical Investigation* from the 1930s, 1940s, and 1950s, it appears to the authors that it was historically more of a common practice to publish raw data as part of an article. There has been a shift in the past 50 years or so from publicly accessible research data to no access unless by personal contact; moreover, most research data are unfortunately discarded either by authors or universities, and then even personal contact is unavailing.

The controversial circumstances of the protective effectiveness of the melanoidins-sunscreen concept resulted in a Grand Rounds lecture on a university's website that discussed the issue in-depth. This Grand Rounds presentation is an open access publication, which can be reviewed by anyone through the Internet (Fusaro and Rice, Presentation). This type of university website publishing fills a niche for unlimited in-depth, detailed communications of all university research or controversial concepts

and differs from the few new journals that publish only negative results (Begley).

Given the explosive growth of research data generated in the past 50 years, it would seem important that universities invest money, time and talent in digital archiving and indexing of all raw research data produced within their scholarly communities. Although the authors recognize the establishment of an Institutional Repository could require the institution to incur ongoing financial burdens for staffing, equipment, and preservation, much of this effort could result in a savings by centralizing functions (Gibbons). What is called for is for universities across the country to focus on creating Institutional Repositories linked through a common metadata and search engine (e.g. Google Scholar). This could be the solution for providing local controlled access to all details of published and unpublished experiments generated in their academic institutions. As Singarella points out, the IR can "build on a growing grassroots faculty practice of self-posting research online" (Singarella 20). Libraries are the logical administrative entity to carry out this function. Indeed, it is expertise in material submission, metadata application, access control, discovery, distribution, and preservation that library staff can offer. Only the library can claim expertise in all of these core functions (Gibbons).

Our examples of new conclusions obtained from reviewing the archival data of published reports illustrate the value of digital archiving of all raw research performed at universities; moreover, this new information needs to be disseminated without any delay. If new conclusions are in conflict with accepted dogma and rejected for publication in standard journals, the use of a network of university Institutional Repositories offers a global alternative medium to stimulate further research, and to promote acceptance of new findings by others (Fusaro and Rice, Presentation). Universities exist not only to educate students but also to create and disseminate knowledge for the betterment of humanity. If that knowledge resource is not accessible, if no one can find it or if no one knows where it is or that it even exists, of what use is it?

Works Cited

Begley, S. "New Journals Bet 'Negative Results' Save Time and Money." Wall Street Journal 16 Sept 2006 : B1.

Crow, R. "The Case for Institutional Repositories: A SPARC Position Paper." ARL Bimonthly Report Aug. 223 (2002). 23 Mar. 2007. <http://www.arl.org/resources/pubs/br/2002br.shtml.>.

Davis, P. "Institutional Repositories: Evaluating Reasons for Non-use of Cornell University's Installation of DSpace." D-Lib Magazine Mar/Apr 13.3/4 (2007). 23 Mar. 2007. <http://www.dlib.org/dlib/march07/davis/03davis.html>.

Fitzpatrick, T.B., and M.A. Pathak. "Chemical Binding of Protective Agents to Human Stratum Corneum." Contract #DA49—092-ARO-85. Arlington, VA: Department of United States Army, Army Research Office, 1968.

Fusaro, R. M. Patent 3,920,808. 18 Nov. 1975.
Fusaro, R. M., and E. G. Rice. "The Maillard Reaction for Sunlight Protection." Annals of the New York Academy of Sciences 1043 (2005): 174-83.

---. Presentation. "Keratin Bound Sunscreens." 2004. 27 Sept. 2006. <http://newadonis.creighton.edu/HSL/grandrounds/GrandRounds-2004-October.html>.

Gibbons, Susan. "Establishing an Institutional Repository." Library Technology Report 40.4 (2004): 1-67.

Harnad, S. "Overture: The Subversive Proposal." Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing. An Internet Discussion About Scientific and Scholarly Journals and their Future. Eds. A.S. Okerson and J.J. O'Donnell. Washington, DC: Association of Research Libraries, 1994. 27 Mar. 2007 <http://www.arl.org/sc/subversive/i-overture-the-subversive-proposal.shtml>.

Hollis, B. W., et al. "25-Hydroxylation of Vitamin D3: Relationship to Circulating Vitamin D3." Journal of Clinical Endocrinology and Metabolism In Press.

Hollis, B. W., et al. "Circulating Vitamin D3 and 25-Hydroxyvitamin D in Humans: An Important Tool to Define Adequate Nutritional Vitamin D Status." Journal of Steroid Biochemistry and Molecular Biology 103.3-5 (2007): 631-4.

Lynch, C. A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." ARL Bimonthly Report Feb 226 (2003). 23 Mar. 2007. <http://www.arl.org/resources/pubs/br/br226/br226ir.shtml>.

Matters, M. D., et al. "Developing Metadata to Organize Public Health Dataets." AMIA Annual Symposium Proceedings 1047 (2005). CD-ROM.

Pathak, M. A., T. B. Fitzpatrick, and E. Frenk. "Evaluation of Topical Agents that Prevent Sunburn--Superiority of Para-Aminobenzoic Acid and its Ester in Ethyl Alcohol." New England Journal of Medicine 280.26 (1969): 1459-63.

Pathak, M. A. "Sunscreens: Topical and Systemic Approaches for Protection of Human Skin Against Harmful Effects of Solar Radiation." Journal of the American Academy of Dermatology 7.3 (1982): 285-312.

Rice, E. G. "Dihydroxyacetone Naphthoquinone Protection Against Photosensitivity." Dermatologica 153.1 (1976): 38-43.

Singarella, T. Online Report. "An Overview of Institutional Repositories." Association of Academic Health Sciences Libraries Charting the Future Committee. Mar. 4 (2005). 23 Mar. 2007.