Academic Distribution of PubMed Citations: Implications for Institutional Repositories

Ed Sperr (sperre@neco.edu)

New England College of Optometry, Boston, MA

Abstract

Institutional repositories mark one potential path to making open access a reality. The hope of many advocates is that if enough institutions construct these repositories a "tipping point" will be reached, as a substantial portion of the current research literature becomes publicly available.

If authors are concentrated in a small number of institutions, then few repositories will be needed to capture their research. However, if researchers are scattered among a large number of institutions, a correspondingly large number of repositories would be needed to be effective.

A survey of the last five years of the PubMed database was undertaken in an attempt to gauge the institutional distribution of authors in the biomedical sciences. Affiliation fields were extracted from the records and location information was analyzed. In MEDLINE records, approximately 47% of first authors who work in the United States are affiliated with schools that belong to the Association of Research Libraries. Therefore it is likely that the establishment of a relatively small number of institutional repositories (fewer than 130) could have a significant impact on the current system of scholarly publishing.

Methods

Using the Entrez Programming Utilities[1] from NCBI, a Perl program was constructed to retrieve PubMed records in XML format. The affiliation information from each record containing an affiliation field was parsed using the XML::Twig module[2]. Perl DBI was then used to enter the affiliation information into a Microsoft Access® database.

In this database, affiliations were compared with a list of ARL libraries as well as a list of countries. While the quality of the affiliation information varied from record to record, there was sufficient information to determine a location in most cases. String matching was used to apply institution and country attributes (Fig. 1).

A spreadsheet program was then used to analyze the resulting data.

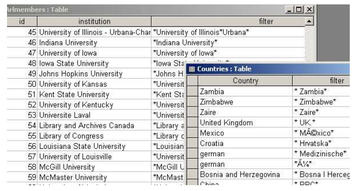

Figure 1: String Matching
Records were associated with institutions and countries by matching the text of their affiliation fields against other tables in the database.

Discussion

PubMed is perhaps the premiere database of the biomedical literature, and as such is widely used by librarians and end-users alike. While PubMed does not index everything, its reach is vast; it covers over 4600 journals[3] and contains over 15 million citations[4]. PubMed is particularly useful for the purposes of this project because it collects information about the institutional affiliation of authors. It is also possible to extract data from the database in an automated way[1].

Institutional repositories are gaining prominence as a way for Academic institutions to both store and showcase their intellectual output. These repositories also become a vehicle for advancing Open Access, as scholars deposit either pre-print or (where possible) as-published versions of journal articles. The hope is that these repositories would eventually make publicly available the bulk of the scholarly production of the institution. The adoption of institutional repositories is still in its early stages however, so it is unclear that this is yet having much impact upon the serials crisis.

While it seems clear that there is some demand for Open Access to scholarly articles, the supply is less certain. There are thousands of institutions of higher education, yet only 363 organizations are currently making an institutional repository available for public harvesting[5]. Although this is a small number in comparison to the total number of institutions generating scholarly publications, my data indicates that even a small number of repositories could have a large effect.

Indeed, the adoption of a repository by only 123 institutions could potentially make nearly *half* of the biomedical articles written in the United States each year available through Open Access.

While there are many challenges to establishing an institutional repository, software packages to implement them are available[6], and a growing user base is starting to share ideas. As the potential benefits are so great, collective organizations such as the ARL should play a role in supporting a wide scale adoption of institutional repositories.

Results

During the five year period of 1999-2003, approximately 85% of all PubMed records contained an affiliation field providing information about the first author of the article. Of these, approximately 17% corresponded to an ARL institution. The picture in even more dramatic when looking solely at citations from the United States. Authors at ARL institutions were responsible for 47% of these citations.

Figure 2: Affiliations of authors in PubMed
Affiliations for 5 years of PubMed citations were analyzed with the purpose of defining how many articles were authored in the United States and how many authors were based at ARL institutions.

References

1. Entrez Utilities [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); [updated: 2004 Aug 9; cited 2004 Nov 30]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

2. xmltwig.com [Internet]. Michel Rodriguez; c1998-2004 [cited 2004 Nov 30] . Available from: http://www.xmltwig.com/

3. NLM technical Bulletin, Jul-Aug, Technical Notes [Internet]. Bethesda (MD): National Library of Medicine (US); [updated: 18 Sep 2002; cited 2004 Dec 1]. Available from: http://www.nlm.nih.gov/pubs/techbull/ja04/ja04_technote.html#milestone

4. MEDLINE Fact Sheet [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2004 Dec 1]. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html

5. OAIster Home [Internet]. Ann Arbor (MI): University of Michigan Digital Library Production Service; [updated: 2 Nov 2004; cited 2004 Dec 2]. Available from: http://www.oaister.org/

6. Budapest Open Access Initiative – A Guide to Institutional Repository Software v. 3.0 [Internet]. [cited 2004 Dec 2]. Available from: http://www.soros.org/openaccess/software/