Integrating the "Green" and "Gold" road to Open Access: Experience from Bioline International

Leslie Chan[1]
Sidnei de Sousa[2]
Jen Sweezie[1]

[1]University of Toronto at Scarborough
1265 Military Trail, Toronto, Ontario Canada
chan@utsc.utoronto.ca

[2]Centro de Referência em Informação Ambiental
Av. Romeu Tórtima, 388, Barão Geraldo
13084-791 Campinas SP Brazil

Abstract

This paper reports on Bioline International's (BI) technical infrastructure, workflow, and hybrid strategy of combing Open Access Publishing (Gold) and Open Access Archiving (Green) to improve the access and visibility of published research from developing countries. It further shows how BI uses open source technologies (e.g. XML, Perl, OAI-PMH) and software (Eprints and DSpace) to promote the widest distribution and discovery of research information for the benefit of international scientific exchange. Data on the general usage pattern of materials on BI are presented. Using the Journal of Postgraduate Medicine as a case study, the effects of open access in terms of improved visibility, citation and author submission are further illustrated. In the process, this paper demonstrates how BI has evolved in response to the changes in technologies and resultant opportunities.

1 Introduction

Few topics in scholarly communication and publishing today are as hotly debated and contested as open access, the idea that scholarly literature should be available online to readers without cost and without most of the permission barriers associated with the traditional pay subscription model of access (Suber, 2003). The primary goal of open access is to improve the visibility and subsequent impact of research publications, generally measured in terms of post-publication citation (Harnad *et al.*, 2004). Indeed some emerging studies strongly suggest a direct correlation between open access literature and their higher citation impact relative to publications that are kept behind toll access (see bibliography on the Effect of Open Access on Citation Impact, compiled by Steve Hitchcock http://opcit.eprints.org/oacitation-biblio.html).

Proponents of open access largely follow the Budapest Open Access Initiative's two prong but complementary strategies of the so-called "gold" road, or open access publishing, and the "green" road, or open access archiving of published research. However, there is considerable debate about the business models for sustaining open access publishing and whether the "green" or the "gold" road is the most effective and immediate route to achieving universal open access. Amid these discussions, there is also increasing discussion about the impact of open access in broadening access to research journals published in developing countries as well as access to research by scholars in the developing world (Chan & Costa, 2005; Chan & Kirsop, 2001). Pioneer open access project such as SciELO (Science Electronic Library Online) of Brazil, coordinated by Bireme and Fundação de Amparo à Pesquisa do Estado de São Paulo – Fapesp (São Paulo State Research Support Foundation), has clearly demonstrated the importance of open access in increasingly the visibility of both Brazilian journals and the citation rate of authors published in "local" or "regional" journals (Coura & Willcox, 2003).

In light of the debates about the relative merits of OA publishing and OA archiving, this paper addresses the pragmatic questions of how best to position neglected journals from developing countries to capture the benefits of open access. Drawing on the usage and download patterns of open access material on the Bioline International web server and eprints archive, we analyze the usage impact of open access publications on the BI system. By further focusing on the *Journal of Postgraduate Medicine* from India as a case study, we highlight the effect of OA on readership and author submissions, which are other proximate measures of the impact of OA. We conclude with observations on the challenges and opportunities presented by OA to high quality "regional" journals and present reasons why funding and development agencies should support OA as part of the overall development strategy for building local research capacity.

2 Bioline International and developing countries journals

Bioline International was established in 1993 to take advantage of the emerging networking technology for online publication, with the intent of lowering the barriers of access to journals from the developing world (Canhos *et al.*, 2001). BI is not a publisher but an electronic aggregator that provides free technical infrastructure and acts as an Open Archive Initiative (OAI) Data Provider. Because most not-for-profit publishers from the developing world lack the financial, human and technical resources for online publishing, BI provides free technical platform and document conversion services to qualified journals that wish to provide free online access to readers. Since 2001, the project became a collaboration between the University of Toronto in Canada, the Centro de Referência em Informação Ambiental, CRIA, in Brazil, and the Electronic Publishing Trust for Development in the UK. Currently there are 31 active journals on the BI system, and many of the participating journals continue to charge for print subscription, although circulation of these journals are typically very small, with few international subscribers.

2.1 Enhancing accessibility and visibility

The BI platform consists of two key servers; the primary server resides in CRIA in Brazil, www.bioline.org.br, while an Eprints server, http://bioline.utsc.utoronto.ca, which mirrors the content of the primary server and stores the files in PDF format, resides at the U of Toronto in Canada. The two servers cross-linked with each other to maximize exposure of the metadata to multiple search services and directories.

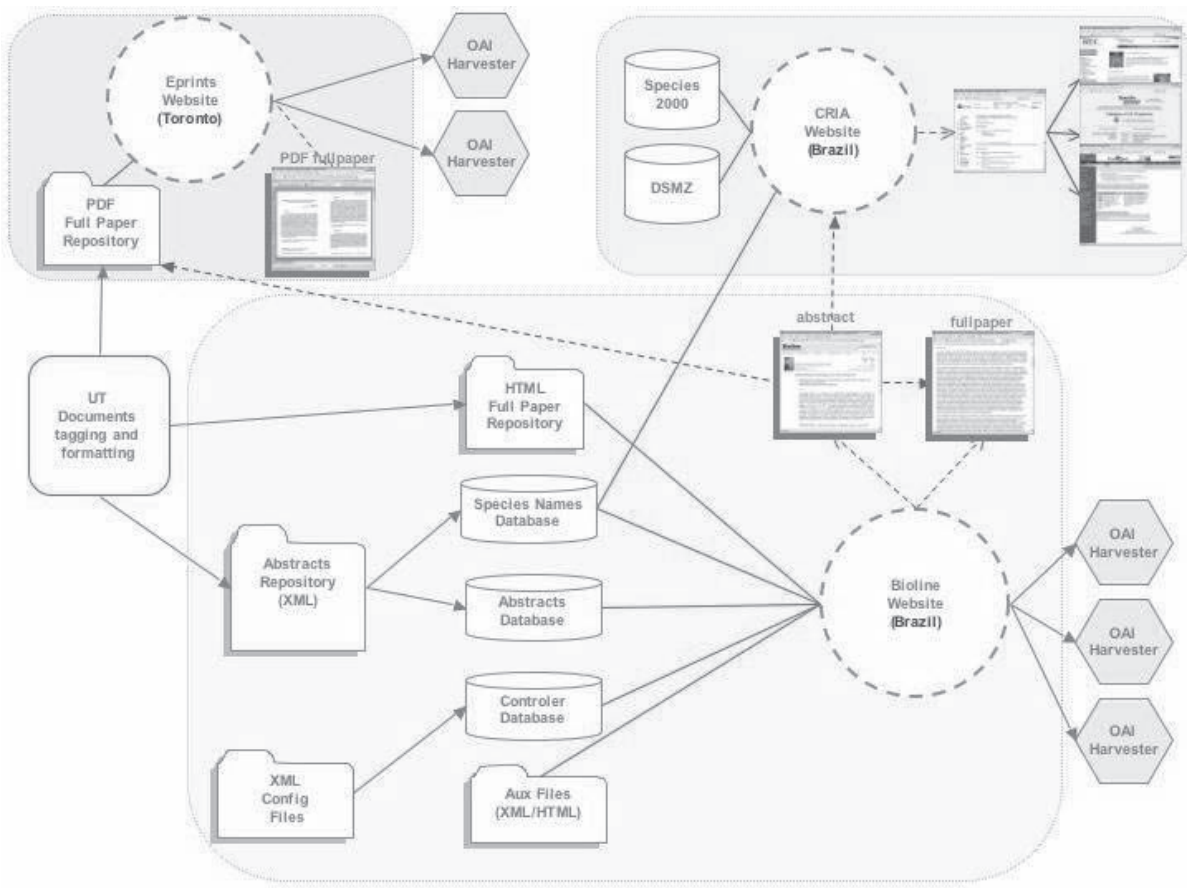

Fig.1 The technical infrastructure of Bioline International, showing the relationship between the Eprints server in Toronto and the primary server in CRIA, Brazil

The workflow
The online publishing process begins when journal publishers email or FTP their press ready files to the BI office in Toronto. Digital files, which arrive in a variety of formats, including Word, Pagemaker, PDF, RTF and HTML, are reformatted so that they can be stored as Bioline specific HTML and PDF documents. During this process, the images, tables, and abstracts, often available in other languages as well as in English, are extracted from the main texts and properly tagged according to a specifically designed XML Schema, producing XML documents that enable better content sharing.

Special attention is given to the species names that are cited within the documents. All species names cited in the text are followed by the special icon  that holds a link to a species search page. The species search page was created by CRIA as a cross reference to the information stored in all systems maintained at CRIA, using species names as a key. By clicking on the icon, the user can immediately have access to all information related to that species available at CRIA (SinBiota, speciesLink, Biota Neotropica, Image Database, etc. and Bioline International itself). The search engine also searches the name in selected external information systems such as PubMed, GenBank, SciELO, ITIS, FishBase.

Receiving, analysing and loading the files
Once the documents are ready, the XML files containing the abstracts, together with the full documents in HTML, are sent by FTP to the main Bioline International server located at the CRIA, where they are stored in the appropriate repositories. The PDF documents are sent to the OAI compliant Eprints server in the Toronto office. The purpose of the Eprints server is to ensure maximal exposure of the journal content through OAI search services. This was particularly important prior to February 2005, when the main BI system was notyet OAI compliant. In addition, the Eprints server is periodically harvested by the digital repository at the University of Toronto's TSpace system to ensure backup, added exposure, and long-term preservation of the publications.

When the server in Brazil receives the documents the updating process is initiated. The process is controlled by a series of Perl scripts that use XML parsers and Database Interfaces packages available as free software. The scripts were designed to require a minimum of human intervention.

The XML files are analysed and, if no errors are detected, the articles are stored in a PostgreSQL database and the tagged species names are stored in a special species names dictionary database maintained by CRIA. In this case, an automatic message is sent to the publishers informing them that the new issues are publicly available.
If the articles present any errors regarding XML tagging, coding etc., a message is sent back to the UT team that takes the necessary action to fix the problems and resend the documents to the server. A controller database is also maintained in the main server holding general information about the publishers and available publications.

OAI Data Provider and commercial directories
Since February 2005, the BI primary server is also available through the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) as an OAI Data Provider. This was achieved by writing specific Perl scripts that are able to understand requests directed to www.bioline.org.br/oai, in accordance with the OAI-PMH version 2.0. Becoming OAI compliant was a major achievement as the material stored is not only available through the main Bioline web site but can be harvested by any other OAI Service Provider existing throughout the Internet.

To ensure maximum visibility of the content, BI is registered with Oaister.org, an OAI Service Provider that harvests a large number of institutional repositories from around the world. Most of the journals on the BI system are listed on the Directory of Open Access Journals www.doaj.org, which also provides metadata for libraries that wish to link directly to the BI journals. Many of the African based journals on the BI system are also indexed and cross-listed in the African Journal Online project www.ajol.org, further increasing the visibility of the contents.

Commercial service providers are also beginning to take note of BI, and EBSCO's A-Z service has been listing BI material since 2003. Since April 2004, the BI Eprints server is being indexed by ISI Web Content, and in March of 2005, Ulrich's Serials Directory also begin listing BI contents. These services promise to improve the visibility of the contents but also integrate the contents into the mainstream databases. We are tracking the outcomes of these inclusions with keen interest.

3 Outcomes thus far

As of April 30th 2005, there are 31 active journals on the BI international site, with a total of 6215 articles. However, the number of articles varies between journals, depending on the time the journal has been on the BI system and on

the age of the journal (see http://www.bioline.org.br/indicadores/conteudo/index?pt). The number of articles that are mirrored on the Eprints server number at 2800 as of the end of April, roughly half the number of articles on the primary server. The unequal numbers are due to the fact that the Eprints server was set up more recently so BI staff are still catching up with mirroring the backfiles on the BI main server.

The ease of accessing and discovering the BI contenthas resulted in a steady increase in the usage of both the PDF and HTML documents. Traffic at both the primary and Eprints site has been growing rapidly, especially in the past year.

Hits and downloads

Total hits for the Eprints server increased from 43,441 in September 2003 to 437,150 in April 2005, while the total number of hits in 2004 amount to 2,014,790, with an average of 167,899 hits per month. The total files downloaded increased from 30,374 in September 2003 to 372,258 in April 2005. The total number of files downloaded in 2004 was 1,218,030 with an average of 95,462 files downloaded per month.

For the primary server, the total number of hits was 114,821 in September 2003 and went up to 966,753 in April 2005. The total hits for 2004 was 3,490,191, with an average of 276,129 per month. Total files downloaded rose from 107,970 to 885,312 in April 2005. Total files downloaded for 2004 was 3,387,812, with an average of 282,317 downloads per month.

	Year		
Journal Name	2002	2003	2004
Memórias do Instituto Oswaldo Cruz (2068 articles)	11526	33001	116971
Neurology India (583 articles)	0	13295	41836
Journal of Postgraduate Medicine (500 articles)	2635	28187	43392
African Crop Science Journal (368 articles)	6319	18556	37716
Agricultura Técnica (263 articles)	856	9946	27621
Indian Journal of Dermatology, Venereology and Leprology (253 articles)	0	1489	14997
Indian Journal of Surgery (237 articles)	0	11256	38389
African Journal of Biotechnology (212 articles)	249	11948	45732
Electronic Journal of Biotechnology (203 articles)	0	9169	37502
Indian Journal of Medical Sciences (157 articles)	0	7894	46358

Table 1. The ten journals with the most number of articles and their respective hits from 2002 to 2004

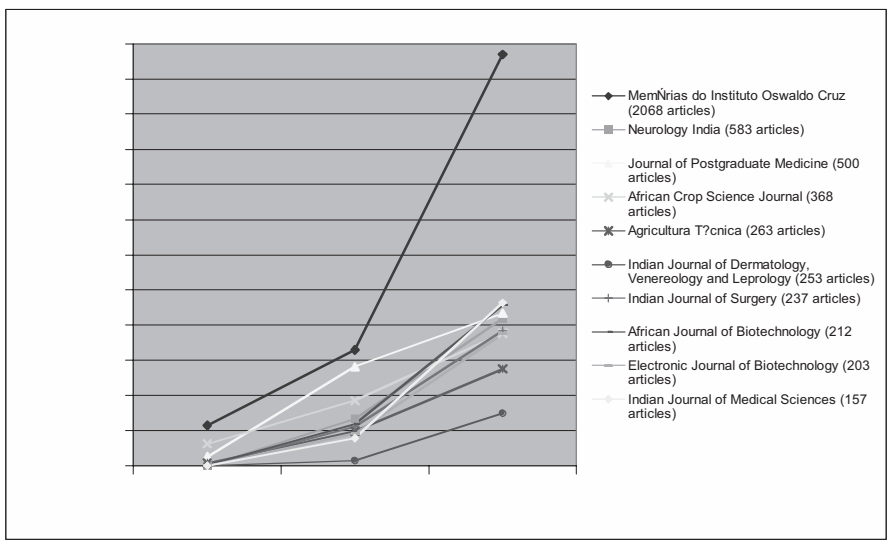

Fig 2. Graph showing the hits of the ten journals with the most number of articles

Referral
Server logs from January 1 to May 1 2005 were analyzed using Funnel Web by Quest Software to determine the pattern of referral, web sites from which users arrive at the BI main server. By far the largest number of referral to the primary server comes from Google (51892), followed by Yahoo (16969). However, the other top ten referral sites are all Google sites from different countries, led by India, Brazil, UK, and Mexico. The DOAJ came in at number 16, with 3295 referrals, while the Eprints server came in at number 22, with 2168 referrals.

For the Eprints server, and for the same period of January 1 to May 1 2005, the top referral site is Yahoo (30347) followed very closely by Google (30192). The Bioline main server in Brazil came in fourth, with 5510 referral. These are likely users who prefer the PDF version of the file they found on the main site. Google India and Canada follow with roughly the same number of referrals (4600). It is interesting to note that no OAI Service Provider was among the top 40 referral site to the Eprints server.

Domains by country
55% (1,631,169) of the traffic to the primary site originates from IP in North America. Unfortunately a large percentage (22.3%) of the IP was not resolved by the server. This is followed by Europe (9.98%), South America (6.11%), Asia (4.13%), Oceania (1.14%) and Africa, which accounts for just over 1% of the total traffic.

For the Eprints server, 53.4% of IP addresses were unresolved. The remaining resolved domains follow the same pattern as the main site, with North America with the most number, followed by Europe, South America, Asia, Oceania, and Africa, the latter again accounts for just over 1% of the total traffic to the site.

3.1 Journal of Postgraduate Medicine (JPGM), Mumbai, India

The JPGM is a publication of Staff Society of GS Medical College and KEM Hospital in Mumbai, India and it was founded in 1955. In June 2002, JPGM joined BI and began to provide free access to its content through the BI platform as well as its own web site www.jpgm.org.

In terms of general usage, JPGM received a total of 2635 hits in 2002 at the BI main server and the hits increased to a total of 43,392 in 2004. In addition to data from server logs, there are other important quality indicators that point to the potential impact of open access. JPGM has in particular documented significant developments in terms of author submissions and the country of origins of contributing authors (Bavdekar & Sahu, 2005).

Increasing Submissions
Total number of submissions increased from 190 in 2000 to 629 in 2004. Number of submissions from authors outside of India has risen from less than 10 % in 2001 to 166 or 38% in 2003 and 189 or 30% in 2004. This is an indication that authors from outside of India are increasingly seeing JPGM as an "international" journal capable of reaching a global audience (data from D.K. Sahu).

Increasing citations
An important consequence of the growing visibility and accessibility of JPGM is increasing citation of its articles, as illustrated in the table below.

Publication year	Citation year	Total number of citations in scientific journals (A)	No of articles other than editorials, letters, and news (B)	A/B
1998-1999	2000	2	60	0.03
1999-2000	2001	12	111	0.11
2000-2001	2002	34	147	0.23
2001-2002	2003	62	155	0.40
2002-2003	2004	137	173	0.78

Table 2: Figures showing the increasing impact factor of JPGM, data by D.K.Sahu (Sources, ISI Web of Science, Scopus, Google Scholar)

Likewise, the citation of articles received in the year of publication has also been steadily increasing.

Year of publication	Number of citations (A)	Number of articles (B)	A/B
2001	0	69	0
2002	3	86	0.035
2003	7	87	0.080
2004	21	74	0.282

Table 3. Journal of Postgraduate Medicine: Citation to articles ratio (2000-2004)

4 Conclusions

One of the guiding philosophies behind BI is the use of simple and tested technologies that reduce the barrier to global online participation. By delivering full text in plain HTML, slow bandwidth users are able to access the content without significant cost and delay. XML encoded abstract enables easy exchange of well defined data, while OAI-PMH further lowers the cost of implementing robust information description, sharing and discovery. Development costs are further reduced with the use of open source software developed by the scientific and library communities.

The data presented here suggests that usage of publications on the BI platform have indeed increased steadily over the past few years and that increased awareness and accessibility of the journals, at least in the case of JPGM, are translated into improved quality of the journals through higher number of submissions, submissions from international authors, and increased citation rate. Although the Impact Factor of JPGM is still very low relative to other main-stream journals indexed in ISI, what is of note is the steady increase in the number over the past few years. What remains to be seen is whether this phenomenon is spreading to the other journals on the BI system, and we are working closely with the various publishing partners to track the necessary data.

What is particularly encouraging is that citation analysis, once dominated by ISI's Journal Citation Report and its Impact Factor, is now giving way to other forms of web based citation linking and impact analysis made possible by open access to the primary research publications. The high referral rates to BI contents by Google and Yahoo indicate that most users are accessing the BI material through free search engines rather than through fee-based services. The new Google Scholar service has the potential to provide even more promising citation linking information for open access quality scientific and academic materials, further enhancing the value of OA (Myhill, 2005).

Developing countries' journals, traditionally excluded by the mainstream indexing services, now have the opportunity to join the global knowledge commons through open access. We therefore urge development agencies and funding bodies who support research publications in developing countries to strongly consider the importance of supporting OA, either through support of OA publishing or OA archiving, or through collaborative partnership with organizations such as SciELO and Bioline International. However, we note from the low access rate of the BI material from regions such as Africa that awareness about OA and its benefits are still limited in many regions of the world. While these low access rates may reflect the low level of Internet access, much work in awareness raising, research, and data gathering on the impact of OA remains to be done.

References

Bavdekar, S., & Sahu, D. K. (2005). Path of progress: Report of an eventful year. *Journal of Postgraduate Medicine, 51*(1), 5-8.

Canhos, V., Chan, L., & Kirsop, B. (2001). Bioline publications: How its evolution has mirrored the growth of the internet. *Learned Publishing, 14*(1), 41-48.

Chan, L., & Costa, S. (2005). Participation in the global knowledge commons: Challenges and opportunities for research dissemination in developing countries. *New Library World, 106*(1210/1211), 141-163.

Chan, L., & Kirsop, B. (2001). Open archiving opportunities for developing countries: Towards equitable distribution of global knowledge. *Ariadne, 30*.

Coura, J. R., & Willcox, L. d. C. (2003). Impact factor, scientific production, and quality of brazilian medical journals. *Mem. Inst. Oswaldo Cruz, 98*(3), 293-297.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., et al. (2004). The access /impact problem and the green and gold roads to open access. *Serials Review, 30* (4), 310-314.

Myhill, M. (2005). Google scholar. *The Charleston Advisor, 6*(4).

Suber, P. (2003). Removing the barriers to research: An introduction to open access for librarians. *College & Research Libraries News, 64*, 92-94.