ELECTRONIC LIBRARIES IN PARTNERSHIP: BEEP FOR AFRICA

By Pier Luigi Rossi
IRD (Research Institute for Development), Bondy, France
rossi@ird.fr

Introduction

Between September 2008 and September 2009, IRD (Research Institute for Development)[1] established a programme of scanning workshops within the SIST project (System for Scientific and Technical Information)[2] of MAEE (French Ministry of Foreign and European Affairs)[3]. This initiative involved documentation centers and libraries of public institutions (Universities, Research Institutes) in several French-speaking African countries (Benin, Burkina Faso, Madagascar, Niger, Senegal, Tunisia).

The programme enabled the purchase of scanning equipment (fast A4 scanners, PCs, dedicated software) and training teams to digitise documents (theses, articles, books) produced by the staff (researchers, engineers, students) of participating institutions. The Greenstone software[4] was chosen to provide access, over the Internet, to the collections of digital materials. All teams were trained in the installation, administration and use of this tool.

The collections made by the project participants consist of documents in pdf format. Metadata is made available either by incorporating existing bibliographic databases (usually in the CDS/ISIS[5] format) or by direct input into the metadata fields of digital files. In the latter case, the metadata entries include document title, authors, publication date and subject.

The project scope was to digitise, enhance and make accessible on the Internet the scientific works of a wide range of institutions from developing countries. Concretely, this initiative organised 30 digitisation workshops in six countries. About 100 people were trained in the methodologies of scanning and creating digital libraries using the Greenstone software. The overall cost of this project (equipment, missions, training, expertise) was 145,000 euros.

Electronic libraries and internet servers: BEEP for Africa

The SIST project also funded the establishment of Internet servers (one server for each country) for sharing scientific information and hosting electronic libraries created with the Greenstone software. These computers were funded by another component of the SIST project. The operability of these countries' SIST servers often met with structural problems that limited their connectivity: low bandwidth, server downtime, frequent power outages, a complex

implementation of IT projects.

In an attempt to provide concrete answers to these difficulties, the IRD installed the BEEP (Bibliothèques électroniques en partenariat) server. This server is located in France, on Bondy IRD site. Its url is www.beep.ird.fr.

BEEP offers several SIST partner institutions temporary hosting of their electronic document collections. This offer is maintained until permanent solutions can be found locally.

This is a cooperative approach that brings together the IRD and partners in developing countries. These institutions desire to share their publications, making them quickly accessible to the scientific communities of the internet world.

Collections can be built by each partner with Greenstone on a desktop computer, by a local administrator trained by the project. The file system generated by Greenstone is then transferred, without any modification, to the administrator of BEEP. In other cases, the IRD is involved in the application design: pdf file collections and logos are transferred to the administrator of BEEP and the collection is built in Bondy. The server uses a Linux version of Greenstone, but collections are created on Windows computers. Greenstone collections can be built in a specific environment but operate and be made available on all systems for which a version of the software exists.

BEEP currently hosts the following collections:

Reports of the National land management project (PNGT2), Burkina Faso (37 documents).

- Dissertations and theses of the Polytechnic University, Bobo-Dioulasso, Burkina Faso (455 documents).
- Reports of the Ministry of Agriculture, Burkina Faso (26 documents).
- Database of MSIRI (Mauritius Sugar Industry Research Institute) publications, Mauritius[6].
- Dissertations of the Interuniversity postgraduate program in Economics, Dakar, Senegal (103 documents).
- Theses and dissertations of the Interstate School of Veterinary Science and Medicine, Dakar, Senegal (498 documents).
- Theses and dissertations of the University Gaston Berger, Saint Louis, Senegal (63 documents).
- Dissertations of INSEPS Institute (Higher National Institute of Education and Sport), Dakar, Senegal (152 documents).
- Dissertations of Polytechnic School, Thies, Senegal (663 documents).

The site also hosts a document collection related to erosion problems in developing countries. This contains 680 articles from an informal journal edited by IRD.

BEEP provides value-added services to all its partners. Statistical access analysis to the contents of the collections (pdf files) can yield strategic indicators on demand. The development of effective tools to achieve better indexing by major search engines and the deployment of OAI gateways provide high visibility to content that is hosted.

IRD feeds regular exchanges with the Greenstone development team (University of Waikato, New Zealand) to participate in the software development. The implementation server under Linux, contributes to developing the skills of computer scientists of the IRD regarding the deployment and optimisation of this solution. The knowledge gained is shared with colleagues who participated in the SIST project, particularly those whose collections are hosted on BEEP.

Search engine optimisation and access analysis

To optimise the indexing of BEEP hosted documents, maintenance and accessibility of a sitemap file[7] are provided. This file contains all the urls of pdf files of each digital collection (Fig. 1). This is a text file generated with a very simple script ("find" and "send" commands)[8].

```
http://www.beep.ird.fr/collect/ptci/index/assoc/HASH0186.dir/2004-Ngom-Politique monaitaire.pdf
http://www.beep.ird.fr/collect/ptci/index/assoc/HASH01a7.dir/2004-Sidikou-Evaluation economique.pdf
http://www.beep.ird.fr/collect/ptci/index/assoc/HASH013e.dir/2002-Dedehouanou-Le franc CFA.pdf
http://www.beep.ird.fr/collect/ptci/index/assoc/HASH0136.dir/2003-Ka-Impact de la liberalisation.pdf
http://www.beep.ird.fr/collect/ptci/index/assoc/HASH11c4.dir/2002-Baguidayem-Capacite contributive.pdf
```

Fig. 1: First lines of BEEP sitemap text file

The site has been declared on several search engines (google, google scholar, yahoo, msn-bing, yopdf, voila, etc.) to better position the collections on the web.

Analysis of access to pdf files is based on the use of log files generated by the BEEP Apache server [9]. The extractions are performed exclusively on pdf files to which an access yields the code 200 (request successfully processed)[10].
A script in php removes the lines generated by the spams [11].

The data are then segmented to create robot access files (including the robot of Google) and users access files. This applies to all pdf files on the server. The same segmentation principles are applied to the log files of each collection.

A comparison between the analysis of log files of documents indexed by Google and the number of files in each collection offers the possibility to verify whether all documents have been indexed. This approach can also show the

number of times each file has been indexed by robots (Googlebot in particular). Figure 2 shows a sample of these results for the collection of the Polytechnic University of Bobo-Dioulasso.

Goobot freq.	File
12	/upb/MSU-1994-BEL-POR.pdf
11	/upb/IDR-2007-YOU-IMP.pdf
10	/upb/IDR-1983-NEK-TES.pdf
9	/upb/IDR-1998-SAV-CON.pdf
8	/upb/ENS-1996-SOM-NEC.pdf
7	/upb/IDR-2008-BAR-ETU.pdf

Fig. 2: Googlebot file extraction scores

Resolving IP addresses for each user access indicates the geographical distribution of visits for the entire site and for each collection. To carry out these treatments, a php script that incorporates the MaxMind GeoIP Country Database[12] was developed.

It was found that 59% of accesses came from Africa. For this continent, the high-usage countries are Algeria (12%), Senegal (11%), Morocco (7%) and Tunisia (7%). For Europe, France represents 26% of the access requests. Haiti (0.3%) is the first country in Latin America.

For the month of March 2011[13] consultations of pdf files (code 200 log files) were 21,308 for the 2651 files available, and 2109 files were really accessed. Thus in March 2011, each file was viewed, on average, 10.10 times and 79.6% of the available files were accessed at least once. The average number of files accessed per day was 687. On average, each day, 26% of the available files were accessed. The same analyses are conducted specifically for each collection of pdf files on the BEEP server. Statistics on the collection of dissertations of the Dakar postgraduate programme in Economics, for example, show that 25% of accesses are from computers with an IP address based in Senegal.

To compare these results, access to documents in pdf format in the IRD legacy collection[14] was obtained for the same month (March 2011). The site provides access to a collection of some 45,000 documents on scientific output (articles, books, reports, theses) of the Institute. For this site, there were 248,777 accesses for 28,844 different pdf files. The average access is 8.62 times. The share of Africa is 29% and the share of France is 30%. For the BEEP server the average access is 10.10 times and the share of Africa is 59%.

Questions asked through Google and which referred to BEEP hosted documents were extracted with a script in php. For each collection, a statistical analysis of words was produced and presented like a tag cloud. The word's size is proportional to its frequency. The tool used to create the word clouds is Wordle[15].

Figure 3 presents the results of this analysis for queries retrieving dissertations of the Dakar postgraduate programme in Economics.

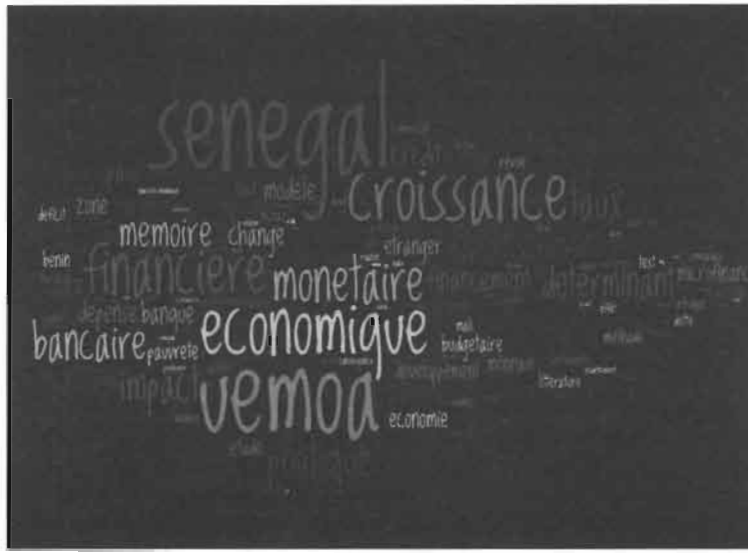

Fig. 3: Tag clouds of the Google queries on dissertations of the Dakar postgraduate program in Economics.

Conclusion

The BEEP (Bibliothèques électroniques en partenariat) server was installed by the IRD, in France, to overcome the structural difficulties that often arise in developing countries for running Internet servers (connections of poor quality, unavailability of servers, frequent power outages, complexities of implementation of IT projects). Presently BEEP hosts 9 digital collections, composed of 2651 files in pdf format. The available documents are dissertations, theses and reports produced by African institutions based in Burkina Faso and Senegal. Given these locations and topics covered, the contents concern primarily, but not exclusively, to the developing countries,

particularly French-speaking African countries. The site provides value added services to institutions who have hosted collections, including statistical access analysis. After several months, although the site is hosted in Bondy (France) analysis shows that about 60% of Internet accesses are from Africa. In March 2011, each file was consulted 10.10 times and 79.6% of available files were accessed at least once.

These access results exceed those of the pdf documents on IRD legacy fund, also hosted on a server based in France.

This analysis shows that effective hosting (overcoming the structural difficulties impeding Internet server implementation in developing countries), combined with careful web positioning, provides high visibility for the documents hosted on the BEEP server. Moreover, the geographical distribution and total number of accesses show that the server location in France does not seem to be a handicap for communities of developing countries. The analyses of accesses generate information on issues and interests of Internet users in relation to the content made available. This kind of information provides strategic guidance for policy makers of Institutions producing documents hosted by BEEP.

The sharing and regular supply of produced statistics, analysis conducted and know-how acquired to the project partners are useful scientific and technical information for the professional community directly involved or interested in the developing countries. The access data are also particularly interesting for the overall debate on the practices and constraints of users in African countries vis-à-vis electronic resources on the Internet.

BEEP can hosts other Greenstone collections created by developing countries institutions needing to improve the visibility and accessibility of their scientific publications.

Bibliography

Andrieu, 0. (2010) *Réussir son référencement web*, Paris: Eyrolles.

Croll, A. and Power, S. (2009) *Complete web monitoring*, Beijing, Farnham: O'Reilly Media.

Rossi, P.L., Ngoma-Mouaya, M. (2000) *"Pleins_Textes"* : *IRD* (Institut de Recherche pour le Développement) electronic library, *International Online Information Meeting*, 24, pp. 201-206.

Witten I.H., Bainbridge D., Nichols, D.M. (2009) *How to build a digital library*, Amsterdam, Boston: Morgan Kauffmann.

Notes

1 http://www.ird.fr/

2 http://www.sist-sciencesdev.net/

3 http://www.diplomatie.gouv.fr/fr/

4 http://www.greenstone.org/

5 http://en.wikipedia.org/wiki/CDS_ISIS

6 This collection does not include documents in PDF. It is not taken into account in this study.

7 http://en.wikipedia.org/wiki/Sitemaps

8 http://en.wikipedia.org/wiki/List_of_Unix_utilities

9 http://httpd.apache.org/

10 http://en.wikipedia.org/wiki/HTTP_code

11 The access to PDF documents have been calculated by removing the recursive accesses. The accesses with the same IP address and the same file must be done with a difference time greater than 5 minutes.

12 http://www.maxmind.com/app/country

13 At the time of writing this article, "March 2011" month was the most recent.

14 http://www.documentation.ird.fr/

15 http://www.wordle.net/