CREATION OF A DIGITAL AFRICAN ARCHIVE

Pierre Malan
Executive Director: Sabinet
pierre@sabinet.co.za

Presented at the 2nd International Conference on African Digital Libraries and Archives (ICADLA-2), University of Witwatersrand, Johannesburg, South Africa, 14th – 18th November, 2011

Abstract

Sabinet Gateway, a non-profit organization that promotes and supports library and information services in Africa, has been awarded a $1,8 million grant from the Carnegie Corporation New York to create an African Online Journal Archive. This archive, the first of its kind to contain purely African content, will make academic inputs from all over Africa available for research purposes to local and international organisations and academic institutions. The aim is to create for the first time a central full-text repository of retrospective journal content that contains important African research across a number of fields, including the medical, social sciences and environmental arenas. These materials have unique value, providing not only the vital groundwork for further or related research but assisting to preserve the heritage of the African continent. Stretching over four years, this project includes the sourcing of African journal content, the negotiation of publisher agreements, digitization and indexing of the journal content and the creation of a front end that will make the journal content easily accessible to end users online. As a result the project aims for the archive to contain approximately 90 000 articles.

Background

Sabinet Online, a public company, currently makes 324 Southern African journals from 155 publishers electronically available. These journals, however, are only made available electronically from the moment that the publisher decides to go this route. The retrospective content is not added to the service.

It has become apparent through the growth and usage of the service that there is significant interest and value in making southern African journals retrospectively available electronically. In 2006 158 000 full text documents were viewed and by 2011 this figure had already increased to over 300 000. It appears that there would also be great value in making a large number of African journals available electronically, as such an archive does not currently exist.

We started with a project to investigate and evaluate a possible African Archive of Electronic Journals for the greater good of the community. Funders were approached and in June 2008 the Carnegie Corporation awarded a grant to Sabinet Gateway of $1,8 million toward digitizing a retrospective collection of 250 journals published in Africa. This paper covers the experiences and lessons learned as well as the progress to date.

The most significant milestones that we reached during this period have been the completion of staff recruitment, the development and launch of the website, and the acquisition and digitization of journals, including metadata creation of articles. There are now 119 journals currently hosted on the site. It was decided to name the project the African Journal Archive (AJA) project.

Research into the project

The creation of an African Electronic Journal Archive was researched and a business and project plan was proposed which addressed the following issues:

- the countries that should be included;
- the journals that should be included;
- how far back the archive should go;
- principles for deciding on content and evidence of demand;
- technology to be used for storage and retrieval;
- a business model and sustainability;
- intellectual property issues and licensing;
- pricing for different countries;
- digitization processes;
- standards; and
- lessons learnt from similar projects.

The research into the project was carried out by two members of staff of the University of Pretoria library, Elsabe Olivier and Monica Hammes, and was based on the following rationale:

Academic authors place a high emphasis on the quality of journals in which they publish. Consequently the strategy for compiling the list of journals was to concentrate on African journals which adhere to the peer review process. This is a process by which all contributions by authors are refereed by scholars recognized as experts in their field of study. The South African Department of Higher Education and Training (DHET) requires South African authors to publish in accredited journals in order to receive subsidy. The Institute of Scientific Information (ISI) comprising Science Citation Index, Social Science Citation Index and Arts and Humanities Citation Index, the International Bibliography of the Social Sciences (IBSS) and the South African DHET accredited list of journals were used as the basis for selecting the African journals. This list also includes journals which meet the accreditation standards but were not accepted by ISI or IBSS on account of their downscaling. The primary goal was not only to identify African journals which could be digitized retrospectively but also to include those that adhere to internationally accepted standards of peer review or accreditation.

The list featured 262 journals published in Africa and covered a variety of subject fields: theology, health, law, education, economics, social sciences and natural sciences. The list also included, for each journal, the ISSN number, publisher's details, journal or publisher's URL (if available), frequency rate, start date, country and subject as well as accreditation status.

The majority of the African journals listed in the ISI and IBSS indexes still come from South Africa and consequently the majority of peer-reviewed and accredited journals are South African. It is widely acknowledged that Africa sources among the lowest numbers of peer-reviewed journals in the world.

The starting dates of each journal were checked either in the South African National Union catalogue (SACat) and WorldCat databases or the publisher's website. Whenever possible the platform which gave access to the electronic full text of the articles was also listed. Journal titles were excluded if they

- were already available totally as online full text journals on the Sabinet SA ePublications service (e.g., The African Finance Journal);
- were published in the United Kingdom or America, e.g., Journal of African History (also a JSTOR title); or
- were published in a language other than English, e.g., French.

We also asked Peter Limb, who is currently Africana Bibliographer/Librarian at Michigan State University, as well as a past chair for the Africana Librarians Council, for his opinion of the value of online access to the journals that have been identified as part of this research. His response was that these appeared to be important journals, though their value would vary from institution to institution and department to department: some users would prefer history, others would prefer sociology and so on. Using the subject discipline of history as an example, the report has included some of the best journals, e.g., South African Historical Journal, for which online access would indeed be useful.

At that time none of the publishers were approached regarding the retrospective digitization project, since it was not certain that this project would go ahead. At this stage we were therefore unable to guarantee that we would be able to secure permission for all these journals from the respective publishers.

There are currently 262 journals that have been identified as part of the research. We would attempt to get permission to digitize these journals retrospectively for at least the past ten years. Depending on the value of the journal, copyright issues and availability of back issues we would attempt to digitize retrospectively to the first issue. It was decided to focus initially on the accredited journals: The South African DHET currently has 242 journals on their accredited list, 143 of which are already in the Sabinet SA ePublications service.

Technology to be used for storage and retrieval

This project will not be creating and maintaining an archive of original print source material, but will only be focused on the provision of a digital archive. The material used to create the digital archive will be returned to the provider of the content. The feasibility and sustainability of the project is directly dependant on the choice of digital collection management software used for ingesting, hosting and accessibility of the digital objects:

access to and use of the digital objects are the required and envisioned end-results of the project.

As part of the research undertaken to determine the viability of the project, time and effort was spent on investigating software options. The research aimed to identify applicability of the software solution in terms of:

1) the extent to which the software/system would support the vision and conceptual strategy;
2) the ability to deploy and use the system within the understood and proven workflow processes (within the organization and in terms of possible partnerships);
3) the extent to which the cost would address issues of investment and fiduciary responsibility;
4) the technical architecture and standards compliance;
5) the extent of support and service that would be required (internationally); and
6) the functionality usability that would be offered.

Sabinet Online currently uses the SiteSearch software to host the Sabinet SA ePublications service. SiteSearch was originally developed by OCLC in the early 1990s and was completely re-written in the late 1990s. In 2001 OCLC stopped further development of the software and released it as open source. The software is still available as open source, but unfortunately the community has not developed the software much further. The software is very stable, but lacks the ability to conform to some of the newer standards and further development of the software is complicated. As a result of this we did not recommend using SiteSearch for the hosting of the archive.

We have also investigated using Open Journal Systems (OJS) which is an open source journal management and publishing system that has been developed by the Public Knowledge Project. The software works exceptionally well as a journal management system as well as for individual journals but during our research we found the following limitations:

- it does not work well for large collections of journals. We experienced response time issues when a large number of journals were loaded into the system;
- it does not handle multiple collections well;
- user authentication management is cumbersome.

The outcome of the research indicates that CONTENTdm® was the preferred choice for digital collection management software. CONTENTdm® is supplied and supported by OCLC and is continually developed in partnership with various research and academic institutions, comprising more than 400 licensed clients and more than 1000 user sites.

Digitization and production processes

We decided to adhere to the JSTOR scanning specifications.

1. The Publisher Liaison Officer will contact the publishers, negotiate the non-exclusive agreements and finalize the publisher requirements for digitization. Once the agreement

is in place then the content will need to be sourced, either from the publisher or from a nearby library that holds the journal.
2. Logs will be kept to track the issues that have been sourced and digitized.
3. When the issues have been received they will be prepared for scanning.
4. The issues will be scanned and imported into CONTENTdm® as images.
5. The metadata for the articles will also be created when the article is uploaded to the CONTENTdm® server.

Standards, intellectual property issues and licensing

Sabinet Online and JSTOR have held discussions and both parties have agreed to work together with regard to the digitization of African journal content to ensure that there is no duplication of effort and that any efforts by Sabinet Online and JSTOR are complementary.

Sabinet Online also held a conference call with Jason Phillips and Kimberly Lutz, JSTOR Director of Publisher Relations. During this call it was agreed that the African Journal Archive would use the same metadata and data standards as JSTOR to ensure forward compatibility with JSTOR. JSTOR also indicated that they would be interested in providing linking access to African journals in the archive. It was further decided that, where practical, the same basic publisher and user agreements would be used for the archive, to minimize confusion for publishers and end users. These agreements would cover use and intellectual property issues. Non-exclusive agreements will be negotiated with the publishers.

Some of the key points of the publisher agreements are:

- The agreement will be not exclusive.
- Publishers will represent rights in the journal issues as "collective works" (that is, we will ask them not to assert copyright in individual articles).
- Sabinet Gateway will have the final decision making power over whether to remove anything from the archive. If a publisher requests that an article or image be removed, Sabinet Gateway will agree to consider the request, but in the end, the decision will be made by Sabinet Gateway.
- Once a journal is in the archive, it will not be removed. If a publisher cancels the agreement, we will agree to not license the journal to new libraries and we will not add new issues. All libraries with access to the title before cancellation will retain access.

We would also like to investigate further the possibility of the archive being included in Portico in future. We have unfortunately not been able to fully investigate this possibility as part of the planning project.

Business model and sustainability

At the inception of the project we were planning to use the "moving wall" principle for the journals. The "moving wall" represents the time period between the last issue available in the archive and the most recently published issue of a journal. Since we had not yet made contact with the publishers at the planning stage, we had not yet decided whether we would have a standard moving wall (e.g. 5 years) for all journals or whether it would be necessary

for the publisher to be able to select a moving wall period. The content may be available in other resources besides the archive: for example Sabinet SA ePublications contains some content dating from 2000 but we would include this content in the archive to maintain the moving wall.

Publishers will not incur any costs for having their journals digitized and made available in the archive; they will however receive no income from the subscriptions to the archive. This model is different from other models where there is provision for revenue sharing. The subscriptions to the archive will ensure the sustainability of the archive, but we would like to make subscriptions as affordable as possible to subscribers. The agreement with the publishers will be a non-exclusive agreement.

The income from subscriptions for access to the archive will be used for:

- maintaining the "moving wall" at an estimated average of 2 issues per journal (262 journals) per year;
- adding new journals to the archive at an estimated 5-10 new journals per year;
- hardware and software licenses;
- hardware and software replacement when necessary;
- provision for backup and recovery of the archive in case of any incident; and
- provision for a change in standards and/or technology and the possible migration to new technology as a result of this.

Should a publisher require the journal to be made available as an open access journal then the publisher would be required to pay a once-off archival fee, which would be used to ensure the sustainability of the archive.

Sabinet Online currently has 102 existing subscribers to the SA ePublications service. These subscribers were seen as potential subscribers to the archive as well. Sabinet Online is actively increasing the number of international subscribers through various marketing initiatives, including advertising, word-of-mouth, direct contact and by enhancing the discoverability of the journals through initiatives such as Google Scholar.

The original plan was for the archive to be subscription-based to ensure long term sustainability of the archive. This model was later amended and to date the archive can be accessed at no cost. The current plan is still to find a preferred business model in future to ensure the long-term sustainability of the archive whilst maintaining on an open access (free of charge) platform to users worldwide.

The open access model has received widespread acceptance. Digital rights management remains a constant concern during contractual negotiations. While publishers are clear that the copyright (of the content) remains with the publisher, the assigning of digital rights to Sabinet Gateway often leads to closer scrutiny and negotiations.

Time frames and project plan

It was envisaged that the project would commence on 1 July 2008 and be completed by 30 June 2012. To date a total of 125 journal titles have been published on the African Journal Archive website (http://www.ajarchive.org). In addition to the journals already published, there are six titles in process (either being scanned or indexed) and we are in negotiations with a further 98 publishers.

The full staff complement consists of the project coordinator, seven full-time staff and seven freelance data analysts. There are two divisions, Publisher Liaison and Production (comprising digitization and data analysis).Sabinet management and staff continue to support the AJA project team with regard to aspects that include website design, marketing, systems and IT support.

Publisher liaison

The collection development framework is being refined as the process continues. New titles that are added to the core list of 262 titles are documented: titles are added if they complement existing key disciplines; result from a direct approach by the publisher or leads from the Sabinet SA ePublications service; emerge from the credibility or ranking of the research body or journal accreditation; or a combination of these factors.

Signed contracts are the outcome of publication research, communication with publishers, and following up of new leads. A database has been set up to manage the data, track the status of negotiations and extract statistical analysis.

Sourcing of hard copies has presented more challenges than merely concluding publisher agreements. Preference is for receiving duplicate copies in the original soft cover binding (rather than hard bound) to ensure high productivity: copies are stripped and scanned in automatic feed scanners; the journals need not be rebound and returned but can then be discarded.

Alternatively "non-destructive" scanning is carried out, but is costly and time consuming. The material is borrowed from local libraries or directly from the publisher. Stripping and rebinding of the journals is outsourced.

Production (digitization and data analysis)

Digitization

Digitization specifications have been adopted according to international standards. Both TIFF (preservation standard) and PDF (download standard) formats at 300 DPI are created. The scans are cropped and the articles are extracted from the issues, to be indexed by data analysts. Originals with grey scale, colour, text and combinations of these are scanned in-house. Once all available copies have been digitized, two sets of the data are copied to storage media. One copy of the data is sent to the publisher and the second is stored on the AJA server.

Data analysis and content upload

The digital collection management platform, CONTENTdm®, has been configured and customized for optimal workflow. The workflow has been streamlined for more efficient and faster upload speeds. Articles are indexed by one full-time metadata controller and seven freelance indexers. The metadata controller manages the production process and liaises closely with the Digitization department. Articles are indexed at a rate of 3500 per month and uploads to the website occur daily.

We have also implemented a Handle System for creating persistent linking. Persistent identifiers are unique names for digital objects on the internet and their use ensures that the object will persist over changes in location. This enables libraries to embed them into their own collections.

Launch of the African Journal Archive Website

The African Journal Archive website (www.ajarchive.org) was launched on 26 May 2010. The announcement was sent to various listservs and the AJA contacts lists. The website provides a form for comments and suggestions. The input is checked daily and visitors are encouraged to join the mailing list. Now that a critical mass has been reached, publishers are starting to approach the archive to request participation. Use of the archive has also grown substantially despite no formal marketing campaign.