CREATION OF A DIGITAL AFRICAN ARCHIVE THROUGH COLLABORATION

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

CREATION OF A DIGITAL AFRICAN ARCHIVE THROUGH COLLABORATION

Pierre Malan (pierre@sabinet.co.za)

Executive Director: Sabinet

ABSTRACT

Sabinet facilitates collaboration between various role players in the African region in an effort to create the most comprehensive collection of African research content in the world. By collaborating with publishers, libraries and faculty at institutions, Sabinet is bringing together full-text content published in journals and institutional repositories, on a single and easily accessible platform.

The initiative was started in 1998 with a project to collect and make available African journal content online. The aim was, for the first time, to create a central full-text collection of journal content containing important African research across a number of fields, including medical, social sciences and environmental. In 2007, the project was boosted further when Sabinet was awarded a grant from the Carnegie Corporation of New York to retrospectively digitize and make available an archive of the content of journals that was published prior to the start of the initiative. There are now more than 340 journals available online, containing more than 200 000 articles. Recently, Sabinet has extended collaboration with academic institutions by harvesting and incorporating content published in institutional repositories into the central platform. This collection of unique African research content is now available to local and international organisations. The result of this collaboration has a unique value, providing not only the vital groundwork for further or related research but assisting to preserve the heritage of the African continent whilst at the same time providing valuable world-wide exposure to African publishers.

1. Background

Sabinet Online, a South African based company, is currently making 324 Southern African journals from 155 publishers available electronically. 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 retrospective Southern African Journals available electronically. In 2006, 158,000 full text documents were viewed on the Sabinet platform and in 2012 this figure has already increased to over 400,000.

It would appear that there would also be a great amount of value in making a large number of African Journals available electronically as such an archive does not currently exist.

We then 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 of $1,8 million to Sabinet Gateway 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 183 journals currently hosted on the site. It was decided to call the project the African Journal Archive (AJA) project.

2. 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:

- Which countries should be included
- Which journals should be included
- How far back should the archive go
- Principles for deciding on content and evidence of demand
- Technology to be used for storage and retrieval
- Business model and sustainability
- Intellectual property issues and licensing
- Pricing for different countries
- Digitization processes
- Standards
- Lessons learnt from similar projects

The research into the project was done by two of the University of Pretoria library staff, 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 and consequently the strategy for compiling this list was to concentrate on African journals which adhere to peer review. Peer review is a process by which all contributions by authors are refereed by scholars who are recognized as experts in their field of study. The South African Department of Education requires South African authors to publish in accredited journals in order to receive subsidy. The ISI (Institute of Scientific Information which consists of the Science Citation Index, the Social Science Citation Index and the Arts and Humanities Citation Index), IBSS (International Bibliography of the Social Sciences) and the SA Department of Education's accredited list of journals were used as the basis to select the African journals. This list also includes journals which meet the accreditation standards but were not accepted by ISI and 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 it covered a variety of subjects: theology, health, law, education, economics, social sciences and natural sciences. It also included each journal's ISSN number, the publisher's details, the journal's or publisher's URL (if available), frequency rate, start date, country and subject as well as the 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 a known fact that the level of peer-reviewed African journals is among the lowest in the world.

The journal's start dates were checked either in the South African National Union catalogue (SACat) or 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.

Journals were excluded if they:

- Are already in total available as online full text journals on the Sabinet African Electronic Journal service, e.g. The African Finance Journal.
- Are published in the United Kingdom or America e.g. Journal of African History (also a JSTOR title)
- Were published in languages other than English, e.g. French.

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

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

There were 262 journals identified as part of the research. The plan was that e would attempt to get permission to retrospectively digitize these journals for at least the past 10 years. Depending on the value of the journal, copyright issues, and availability of back issues we would attempt to retrospectively digitize to the first issue. It was decided to in initially focus on the accredited journals.

The South African Department of Education currently has 242 journals on their accredited list of which 143 are already in the Sabinet African Electronic Journal service.

2.2 Technology to be used for storage and retrieval

The material used to create the digital archive will be returned to the provider of the content. This project was only focused on the provision of a digital archive.

The feasibility and sustainability of the project was 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 being the required and envisioned end-result 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 supports the vision and conceptual strategy;
2) The ability to deploy and use of the system within the understood and proven workflow processes (within the organisation and in terms of possible partnerships);
3) The extent to which the cost addresses issues of investment and fiduciary responsibility;
4) The technical architecture and standard compliance;
5) The extent of support and service required (internationally);
6) The functionality offered and usability.

At the time, Sabinet Online used the SiteSearch software to host the Sabinet African Electronic Journal service. SiteSearch was originally developed by OCLC in the early 1990's and was completely re-written in the late 1990's. 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 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.

- does not handle multiple collections well

- user authentication management is cumbersome

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

Since the start of the project Sabinet has migrated all their content to a SOLR based platform, whilst production is still taking place on **CONTENTdm®** it is envisaged that all end user access services will be through the much improved platform.

2.3 Digitization and production processes

We decided to adhere to the JSTOR scanning specifications.

Firstly the Publisher Liaison Officer contacts the publishers and negotiates the non-exclusive agreements and finalizes the publisher requirements for digitization.

Once the agreement is in place then the content is sourced. The content is either sourced from the publisher or from a nearby library holding the journal.

Logs are kept to track which issues have been sourced and digitized.

When the issues are received they are prepared for scanning.

The issues are scanned and imported into **CONTENTdm®** as images.

The metadata for the articles is also created when the article is uploaded to the **CONTENTdm®** server.

2.4 Standards, intellectual property issues and licensing

Sabinet Online and JSTOR have had discussions in which both parties agreed to work together regarding the digitization of African journal content to ensure that there is no duplication of effort but rather that any efforts by Sabinet Online and JSTOR are complementary.

Sabinet Online also had a conference call with Jason Phillips and Kimberly Lutz, JSTOR's Director of Publisher Relations. During this call it was agreed that the African Journal Archive will 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 also decided that where practically possible 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 the use and intellectual property issues.

Non-exclusive agreements were negotiated with the publishers.

Some of the key points of the publisher agreements are the following

- the agreement will be not exclusive
- publishers will represent that they have rights in the journal issues as "collective works" (that is, we will not ask for them 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 their request, but in the end, the decision will be made by Sabinet Gateway.
- Once a journal is in the archive, it cannot be removed. If a publisher cancels their 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 further investigate the possibility of the archive being included into Portico in future. We were unfortunately not able to fully investigate this as part of this planning project.

2.5 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 will 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. The content may be available in other resources besides the archive as well. For example the Sabinet African Electronic Journal service contains some content that is from 2000, but we would also include this content in the archive to maintain the moving wall.

Initially a subscription model was envisaged. Publishers do not incur any costs in having their journals digitized and made available in the archive, and they would receive no income from the subscriptions to the archive. This is a different model than other models, where there is provision for revenue sharing. The subscriptions to the current issues would ensure the sustainability of the archive, but we wanted to make the subscriptions as affordable as possible to subscribers. The agreement with the publishers is a non-exclusive agreement.

The income from the subscriptions for access to the archive would be used for the following

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

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

Sabinet Online currently has 112 existing subscribers to its African Electronic Journal service. These subscribers were seen as potential users to the archive as well. Sabinet Online is actively increasing the number of international subscribers through various marketing initiatives, including amongst others, 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 sustainable 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 the publishers are clear the copyright (of the content) remains with the publisher, the assigning of digital rights to Sabinet Gateway, often leads to closer scrutiny and negotiations.

3. Time Frames and Project Plan

It was envisaged for the project to commence on 1 July 2008 and be completed by 30 June 2012. The project has subsequently been extended for a further year till June 2013. To date a total of 183 journal titles have signed to be published on the African Journal Archive website (http://www.ajarchive.org). Of these titles 123 are already published on the website with the rest of the titles in process (either being scanned or indexed). There are now 108 132 journal articles available on the archive.

The full staff complement consists of the project coordinator, 7 full-time staff and 7 freelance data analysts.

There are two divisions, namely Publisher Liaison and Production (comprising digitization and data analysis

Sabinet management and staff continue to support the AJA project team with regard to aspects including website design, marketing, systems and IT support

4. Publisher Liaison

The collection development framework is 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, leads from the African Electronic Journal service, credibility / ranking of the research body, journal accreditation, or a combination of these factors.

Signed contracts are the result 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 just concluding publisher agreements. The preference is to receive duplicate copies in the original soft cover binding (rather than leather bound) to ensure high productivity. Hard copies are stripped and scanned in automatic feed scanners. The result is the journals can be discarded, not rebound and need not be returned.

Alternatively "non-destructive" scanning is applied. This 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.

5. **Production** (Digitization and Data Analysis)

5.1 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 issues extracted into articles to be indexed by the 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.

5.2 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. The articles are indexed by one full time metadata controller and 7 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.

6. 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 which is checked daily and visitors are encouraged to join the mailing list. With a critical mass having been reached, publishers are starting to approach the archive to request participation, however this has now been stopped with the looming project deadline in sight. Use of the archive has also grown substantially despite of no formal marketing campaign.

7. Conclusion

The African Journal Archive is a unique repository of African Journal content from South Africa, Ghana, Tanzania, Botswana, Nigeria, Zimbabwe, Kenya, Namibia and Ethiopia. This valuable resource would not have been possible if it was not for the initiative and funding from the Carnegie Corporation and the collaboration from all the stakeholders on the continent. In future the Archive will be combined with all current journals hosted by Sabinet which will create a resource of more than 300 000 African full text articles which will be the largest resource of its kind in the world.