Data Archiving in Africa: The South African Experience

by Maseka A. Lesaoana[*]

Introduction

The South African Data Archive (SADA) was established in 1993 by the Centre for Science Development (CSD) of the Human Sciences Research Council (HSRC) in Pretoria. The first staff member of SADA joined the HSRC in January 1994. As a data archive SADA's primary function is to locate, acquire, store and disseminate mainly quantitative machine-readable research data in the humanities and social sciences.

South Africa's past has divided the research community in the country into two major groups: (i) the expert minority comprising mainly advantaged institutions (some universities, technikons and research councils; and the private sector); and (ii) the disadvantaged majority comprising largely universities, technikons and other institutions in the former homelands. In the short term, the expert community will tend to play the role of "depositor" of data at SADA, while both the disadvantaged and expert communities will be users of the data.

Nineteen ninety-four (1994) was an epoch year in the history of South Africa. Apart from political renewal, it heralded a period in which researchers in South Africa are faced with immense challenges in the development of a sound science and technology system. Natural and social scientists are engaged in a number of large-scale local, national, regional and international research projects. However, the bulk of the government's funding generally goes to natural sciences research and its support facilities.

A unique challenge facing South Africa in the transformation process following the first democratic election in 1994 is to implement the National System of Innovation in Science and Technology, as defined in the recently published green paper on science and technology. The paper highlights national data gathering involving large sums of money, and suggests that the data be subjected to public scrutiny. South Africa will soon establish a National Research Foundation (NRF) which shall promote and support research through funding, human resource development, infrastructure provision and capacity building in order to facilitate the creation of knowledge, innovation and development in all fields of science and technology. The NRF draft bill promulgated earlier this year indicates that the CSD of the HSRC will form part of the social sciences and humanities division of the NRF. However, it is not clear whether SADA and other research information facilities will also move to the NRF with the CSD. The future of SADA is shrouded in uncertainty.

Historical background

In February 1993 Per Nielsen, the then Director of the Danish Data Archive (DDA), was invited to the HSRC as a consultant to undertake a two-week feasibility study on the viability of establishing a data archive in South Africa. He wrote a detailed report on his findings, including suggestions based on the experiences of data archives in other countries. Nielsen found the HSRC a suitable location for the archive for the following two reasons: (1) the HSRC has the technical infrastructure needed to improve the quality of research and to provide training at all levels countrywide; and (2) the wealth of research data at the HSRC since its inception in 1969 can be placed at the disposal of SADA for secondary analysis.

Some of the important issues raised in Nielsen's report included the following:

- standards for proper documentation of data
- rationale for establishing a data archive in South Africa
- implementation of the organizational structure
- a variety of financial models including user payments
- recovery of costs
- staff composition
- establishment of an interim advisory body
- types of data at the HSRC that could constitute the first holdings of SADA

Also discussed in Nielsen's report were the existing documentation standards, data-processing issues, international participation and co-operation, and implementation of bilateral co-operation. Nielsen's recommendations were used as a blue-print for the establishment of SADA.

Challenges and developments

Location

SADA seems to be appropriately located at the HSRC, based on the experiences of other data archives, for carrying out its data-archiving activities. Since its inception in 1969 the HSRC has accumulated a wealth of research data, mainly from numerous projects undertaken in collaboration with researchers at various universities countrywide. This provides an opportunity for SADA to liaise with these researchers and their institutions.

At the HSRC, SADA is a division within the Research Information Directorate (RID) of the CSD. Besides housing SADA, RID also maintains a set of databases on human sciences issues through which it provides bibliographical information on current and completed research projects, research and professional organizations, researchers, and forthcoming conferences.

The CSD is committed to redressing the imbalances in research opportunities, and to empowering scholars from historically disadvantaged sectors by actively supporting their participation in the research structures, activities and decision-making processes of the broader research community. It also administers the allocation of various categories of research funding and scholarships to postgraduate students and researchers in the human sciences at South African tertiary institutions and NGOs. The Directorate: Research Capacity Building of the CSD supports the expansion of institutional research capacity by developing research skills among disadvantaged scholars.

Staff

SADA started its operations in 1994 with three staff members who joined the HSRC in January, September and November respectively. Finding suitable staff in a country where data archiving and secondary analysis are new activities is not easy. Furthermore, there are no data libraries or similar facilities in South Africa, and methodology training courses are almost non-existent. The high staff turnover rate at the HSRC has also adversely affected SADA’s progress. Two of the three staff members left the organization (in February 1995 and December 1996). A fourth post of administrative assistant was filled in August 1996.

SADA staff were fortunate to be able to work with experienced data archivists. Repke de Vries of the Steinmetz Archive in the Netherlands visited SADA twice: (1) in October 1994 for a period of five months. The aim of this visit was to offer advice during the initial phase of the development of the archive; and (2) in May 1996 for a period of two weeks in order to study the developments implemented since his previous visit. Shalane Sheley of ICPSR (Inter-university Consortium for Political and Social Research) was contracted to SADA for a period of one year ending in September 1996.

In addition to the four posts originally approved, three new posts were added during 1996. Two posts were filled in January 1997 and one in March 1997, and two are vacant.

While many lessons can be learned through electronic discussions, appropriate technologies to meet the needs of the majority of researchers in a specific country can best be established through interacting directly and exchanging views with the researchers concerned.

The research community

A data archive needs to define clearly the research community for which the services are to be provided, both at supply and demand points. SADA, as a national data archive, has stratified its research community into: (a) the academic community comprising universities, technikons, and training colleges (there are 21 universities and 15 technikons in South Africa); (b) various government departments including the Central Statistical Service, health, education, police, prisons and correctional services; (c) parastatal organizations such as the Human Sciences Research Council (within which SADA is housed), the Medical Research Council, the Social Sciences Development Forum, the Development Bank of Southern Africa, and several economic and social research institutes in the country; (d) research NGOs such as the Community Agency for Social Enquiry (CASE) and the South African Labour and Development Research Unit (SALDRU); and (e) the private or commercial sector, including the mining industry and the Association of Market Research Organizations (AMRO).

By defining the research community, SADA can locate the type of studies undertaken and identify who should receive the archive’s services and what types of data are available or needed. Knowing the research community also assists in the establishment of advisory committees and indicates the target groups to be consulted when assessing the delivery of services by the archive.

Meeting the needs of researchers

Data archives store data in such a way that they meet the needs of the majority of their researchers. Often suppliers (depositors) of research data are also users at the demand point. However, at the demand point new researchers also become involved. An archive’s developmental stages should keep pace with researchers’ interests.

Computer technology plays a major role in data archiving. New developments in computer technology occur regularly, and data archives are faced with the challenge of keeping abreast of these technological advances. The recent switch from the mainframe to the UNIX environment is a typical example. Because of the increasingly large amounts of data involved in research, new distribution and storage mediums and new statistical software packages are emerging. CDROMs and cartridges have gained recognition, and due to their large storage capacity they are employed for storing large datasets in place of the still widely used diskettes. Mainframes' 9-track tapes are being phased out. Data are stored and disseminated in a highly compressed form, and the Internet has brought about new and user-friendly developments in computer technology. The data archivist's job is to meet users' needs effectively, while at the same time ensuring that the data in the archive are preserved for long-term usage.

SADA has devised a user survey form (also available on the Internet) to determine what tools are most often employed by researchers in their analyses. This will assist SADA in catering for its users' needs. To date the response rate has been low, perhaps because in some cases researchers are not yet familiar with computer tools.

Since 1994 SADA has organized workshops at major research centres in the country, mainly at historically disadvantaged universities (HDUs) where the lack of infrastructure has proved to be a major problem. These workshops are aimed at creating awareness of the existence of SADA and its activities.

SADA publishes a newsletter, SADA News, and a draft SADA Guide (catalogue of holdings) to inform the research community of its activities and developments in data archiving. The study descriptions of SADA's collections are also accessible on the Internet (World Wide Web).

Financial and time constraints

Financial constraints impede the establishment of any archive and the supply of data to the archive. The provision of cost-effective services by SADA is not (yet) understood by funders and owners of data in South Africa since secondary analysis itself is not well understood. Even the data that are available are often not properly documented. Data suppliers frequently request funds to enable them clean up their data.

SADA advisory committees

After the visit by Per Nielsen in 1993, an Interim Advisory Committee (IAC) of SADA was set up to offer advice on the establishment of a data archive. The IAC was replaced by the SADA Board, also an advisory body, in February 1996. The board currently consists of 17 elected members representing the research community in South Africa. The SADA Board reports to the HSRC Council. An executive committee of the board was elected in November 1996.

General progress

SADA has progressed well - thanks largely to the help of the established data archives of Europe and the USA. For instance, the above-mentioned visit by the then Director of the DDA, Per Nielsen, contributed greatly to SADA's understanding of data archiving. His investigations into research studies already undertaken in South Africa, guided SADA's acquisition of potential holdings. He also assisted in registering SADA as a member of the International Federation of Data Organizations (IFDO), which led to a visit by the head of SADA to a number of IFDO data archives: the ICPSR in the US, the ESRC Data Archive, the DDA, the Swedish Data Archive (SSD), and the Steinmetz Archive (STAR). The purpose of these visits was to learn how other data archives operate in order to design a suitable model for SADA. It was during these visits that (i) SADA became a member of the ICPSR, and (ii) negotiations took place with Repke de Vries of STAR, resulting in his visiting SADA for a few months to advise the staff at the take-off phase. He gave valuable tips on all aspects of data archiving and data management.

During the planning stage, a CDROM drive, SAS & SPSS, DBMSCOPY and Folio Views were acquired, and the Internet facilities were set in place.

Conclusion

The earliest data archives were established in Europe and North America in the 1960s. Through international associations such as IFDO, IASSIST and CESSDA these archives established guidelines for data archiving and shared their knowledge on the collection, storage, documentation and dissemination of data. Newly established data archives are fortunate to be able to share this knowledge. However, for any data archive general considerations as well as unique considerations pertaining to the specific country have to be taken into account.

General considerations include:

Location: in terms of wealth of and access to research data; institutional credibility; and technical and infrastructural support;

Staff: qualified staff to run the activities of the archive. Initially fewer staff with broad knowledge base are needed;

Networking: networking with researchers and research institutions that produce data is vital. Advisory committees formed with representatives from research institutions can help strengthen networking relationships;

The science of data archiving: through international links such as IFDO, a new data archive can learn the science of data archiving, such as the procedures for acquisition, processing, storage and dissemination of data; and

Funding: this is a major problem. Data archives are not profit-making facilities, and often depend on government funding.

Most specific lessons learned by SADA stem from the economic and political changes currently taking place in South Africa. Data archiving activities are directly linked to the research capacity building activities at higher institutions of learning. Due to apartheid, the education system in the country has been badly skewed. A few privileged institutions had the capacity to undertake research while the majority did not have the resources to do so. Sharing of research between these two groups did not take place. SADA is consequently faced with the challenge of rectifying this situation and also of promoting the new disciplines of data archiving and secondary analysis throughout South Africa and Africa in general.

Other issues to be considered by new data archives:

Research practices: these differ from country to country. Some researchers do not want to give other researchers access to their data for reasons ranging from not being accustomed to the culture of sharing, to wanting to make money by selling their data;

Poor research training programmes: unlike in western countries, South Africa has poor methodology programmes. Summer schools in Northern America and Europe, for example, are held to strengthen the research capacity of data archives and data libraries;

Collaborative research in the region: joint activities by South African researchers are only now starting to take place. Communication breakdown between African countries makes collaborative research difficult. In Europe, for example, the Eurobarometer studies and International Social Survey Programmes are well-known research studies undertaken jointly by European countries.

While some good lessons can be learned from visiting established data archives abroad, the best option is to invite experienced data archivists to the new archive at home. Technology at developed data archives is already at an advanced level and it may be difficult, if not impossible, to implement new practices at the new archive where the infrastructure may be poor or not compatible.

IASSIST and IFDO conferences have traditionally been hosted by Canada, Europe and USA in that order. The mission, objectives and activities of these bodies should be reviewed. One objective should be to draw in more participants, particularly from the third world countries and countries not previously included. Efforts should be made to host such conferences on other continents and to provide funding for participation by researchers in third world countries, thereby increasing membership and building research capacity around the globe.

* Paper presented at the IASSIST/IFDO 1997 Conference, May 6th May 9th Odense, Denmark.