INNOVATIONS FOR THE CURATION AND SHARING OF AFRICAN SOCIAL SURVEY DATA

H L Woolfrey

DataFirst, University of Cape Town, Rondebosch, 7700, Cape Town, South Africa
Email: lynn.woolfrey@uct.ac.za

ABSTRACT

A substantial amount of data is collected through surveys conducted in Africa by national statistics offices, international donor organisations, research institutions, and the private sector. Data management at African national statistics offices is hampered by limited resources. An option for data curation in African countries is the establishment of dedicated institutions for data preservation and dissemination, such as survey data archives, and research data centres. DataFirst, at the University of Cape Town, has established an African data service and is helping to improve African data curation practices through providing data, promoting free curation tools, and undertaking data management training in African countries.

Keywords: Survey data, Social Science data, African data, Data curation, Data preservation, Data sharing

1 INTRODUCTION

A substantial amount of data is collected through surveys conducted in Africa, but only a small percentage of this data is preserved in the long-term, and an even smaller percentage is disseminated as microdata files to support academic research and policy monitoring. Producers of African data include government statistics offices, international donor organisations, foreign and local universities and other research institutions, and private sector institutions. Their data archiving practices vary, from private survey projects that do not share data, to universities and statistics offices that have begun to establish infrastructures for the curation and sharing of their data products.

2 AFRICAN DATA PRODUCERS AND THEIR DATA CURATION PRACTICES

International donor organisations require country-level data to monitor their regional development projects in Africa. To obtain this data they conduct surveys, independently and in partnership with African governments and regional organisations. Donor organisations often make the microdata files from their projects available for further research. For example, the World Bank provides African data via its online microdata catalogue (World Bank, 2011). African universities and other research institutions are also data producers. However, most of these institutions do not have established data preservation or data dissemination policies or practices. Historically, data sharing among researchers in the region has taken place in an ad hoc manner, and as a result, much valuable quantitative research has been unavailable to the wider research community. Despite the advantages of data sharing espoused by the academic community, many African researchers are still reluctant to share their data. This is because the time and resources required to preserve and disseminate data are not available to them. This is also due to some extent to the dissuading motivations inherent in academic research, which is an environment in which exclusive access to original data can give researchers advantages over rivals in an academic field. Surveys conducted by private sector institutions in African countries also collect valuable data. However, although private sector data producers may provide their data for a fee, generally this data is collected for paying clients and is not available for reuse.

National statistics offices are set up by governments to undertake censuses and surveys to collect statistics to provide evidence for government planning. Most data collected and archived in Africa is official data of this nature. Until recently these statistics were used in-house by government statisticians as evidence for national policymaking. Information based on surveys was provided to the public in the form of reports containing tables of aggregated data. However, with growing international emphasis on the importance of statistical data as a national resource for scientific investigation to foster innovation and to provide feedback for sound national decision-making, some African leaders have given support for the preservation and distribution of national

survey microdata for reuse by researchers. Effective government planning is increasingly seen to depend upon sound policy analysis by researchers utilising survey microdata. This enables them to correct or confirm the findings of government statisticians. The role of the research policy interface, in which sound research by academics enables effective government planning, has come to depend on researchers gaining access to original microdata files (Africa Symposium on Statistical Development, 2006; African Union, 2009).

The management and sharing of data by National Statistics Offices is, however, constrained by several obstacles. These organisations have limited financial and staff resources to curate microdata files and ensure their long-term availability. Despite government lip-service to the value of evidence-based policymaking and official claims regarding commitment to harnessing empirical data for economic growth, scant government funding is allocated to statistics offices in African countries. Statistics offices in most countries of the region are chronically underfunded and suffer from shortages of basic equipment, such as computers and vehicles (Woolfrey, 2010). Skills shortages and high staff turnover due to low salaries in the public sector also result in a paucity of analytical expertise in these institutions (Lufumpa & Mouyelo-Katoula, 2005). Government expenditure on statistics in African countries is mainly allocated for data collection, and very little funding is made available to support the long-term preservation and sharing of national data (Kiregyera, 2005). The outcome of this is that in many national statistics offices data curation is not practiced in a systematic manner. This has led at times to data losses or the production of unreliable data (Regional Reference Strategic Framework for Statistical Capacity Building in Africa, 2006).

3 AFRICAN DATA CURATION INSTITUTIONS

Survey data archives and research data centres are dedicated facilities for the sharing of census and social survey data. In Europe a network of national survey data archives fulfils this purpose cross-nationally. These archives provide the advocacy, institutional links, and skilled staff to facilitate data sharing in the region. Data archives acquire, store, and disseminate survey microdata for research purposes (Council for European Social Science Data Archives, 2012). The South African Data Archive, SADA, based in Pretoria, was established in 1996 to formalise the sharing of South Africa's official survey data. Researchers can browse SADA's data portal at http://sada.nrf.ac.za/ahlist.asp and apply online for data to be sent to them on CD or can download the data via an FTP server (South African Data Archive, 2011). SADA is currently the only government funded survey data archive in Africa (Woolfrey, 2010).

Research data centres are university based facilities established to give research communities access to census and survey data. These can be set up as national institutional networks for microdata sharing, for example, the Canadian Research Data Centre Network (CRDCN, 2011). DataFirst, established at the University of Cape Town in South Africa in 2001, has created a research data centre at the university as part of their data service. This research unit also undertakes projects to support data reuse for research and government planning in African countries. DataFirst's work to assist data usage involves a three prong strategy. Firstly, the unit is working to establish itself as a trusted repository for digital research data to ensure data deposits from African data producers. DataFirst utilises innovative technology to make this data easily discoverable and obtainable for research purposes. The unit's research data centre provides data access to students and staff at the university. To extend this access to the international research community DataFirst provides online data access via their website http://www.datafirst.uct.ac.za/catalogue3/index.php/catalog. This site acts as an international portal for African data and for knowledge exchange around data quality issues pertaining to African microdata. Secondly, DataFirst undertakes advocacy work concerning data usage for evidence-based policy-making. For example, staff participates in the data advocacy work of regional bodies such as the UN Economic Commission for Africa.

Thirdly, Advancing data analysis skills among African researchers, the unit's original mandate, is still a key part of its work. This is accomplished by hands-on assistance in the data centre and regular workshops in basic and advanced data analysis. These are well attended by researchers from other African countries. Research recently undertaken by DataFirst, which is in its tenth year of existence, has shown that making microdata accessible and providing the tools to manage and analyse this data increase data demand and advance data quality, to the benefit of research and policymaking in the region (Woolfrey, 2010).

Data quality and data curation best practice are further fostered through the unit's Data Quality Project http://www.datafirst.uct.ac.za/wiki/index.php?title=Category:DataFirst_and_Saldru_Mellon_Data_Quality_Project. The project was initiated in 2006 to investigate the comparability and usability of South African government microdata. Project researchers work with the official data producer, Statistics South Africa, to advance the quality of national data products. The project is aimed specifically at South African data, but the

work includes innovations to support the quality of microdata produced in or about other African countries. The Data Quality Project focuses on improvements to all the quality dimensions of African data to ensure their fitness for use. Of concern are the accessibility, relevance, timeliness, and accuracy of African data as well as its comparability and ease of interpretation (US Census Bureau, Methodology and Standards Council, 2006).

Lessons learned by DataFirst are taken to other African countries through the unit's work with the Accelerated Data Program (ADP). The ADP is funded by the Organisation for Economic Co-operation and Development (OECD) to advance data curation skills to support evidence-based governance in developing countries. The ADP utilises free and open source data curation software to assist governments to preserve, use, and share their national data. The software enables the creation of standardised data descriptions (metadata) to assist data usage and data comparability. The metadata editor is a component of proprietary software developed by NESSTAR, (NESSTAR, 2011) distributed as freeware. Survey metadata created with the editor and data files are shared using web-based software, the National Data Archive (NADA 3.1), created by the Development Data Group of the World Bank. The NADA tool allows the creation of online microdata portals to assist best practice in data curation by resource strapped official data producers in developing countries. The NESSTAR metadata editing software, the NADA software, and guides for their usage are distributed by the International Household Survey Network http://www.surveynetwork.org/home/index.php?q=tools/toolkit (IHSN, 2011). Since 2008 DataFirst has been working with the ADP to install the data curation software at national statistics offices in African countries and provide training in their usage. To date the tools have been adopted by national statistics offices in nineteen African countries. Their optimal use is still hampered by a lack of data sharing permissions and policies in some African countries. However, the availability of these resources has overcome many of the technological barriers to effective data curation in these countries and may set the stage for future cross-country data exchange in the region.

4 CONCLUSION

Evidence from the reuse of social survey data produced by national statistics offices, donor organisations, and research bodies in Africa can provide feedback to governments to support their development planning. Best practice in data curation is not currently followed by African data producers. Data curation by national statistics offices, which are the main producers of survey data in African countries, is frustrated by technological and human resource constraints (Woolfrey, 2010). Most African researchers also do not curate their data products for reuse as there are few rewards in academia for devoting resources to data sharing. In South Africa this issue has been addressed by the establishment of data sharing institutions, such as the South African Data Archive and DataFirst. These data services ensure access to African government microdata otherwise not in the public domain and allow re-examination of data to enhance the role of data as evidence for more effective governance in the region.

5 REFERENCES

Africa Symposium on Statistical Development (ASSD) (2006) Resolutions. In *2006 Africa Symposium on Statistical Development, "The 2010 Round of Population and Housing Censuses"*, Cape Town, South Africa. . Document ASSD2006/03. Retrieved October 29, 2006 from the World Wide Web: http://www.statssa.gov.za/asc/WebsiteReports/ASSD2006-03.pdf

African Union (2009) African Charter on Statistics. Addis Ababa: *Assembly of the African Union*. Retrieved October 10, 2009 from the World Wide Web: http://www.africa-union.org/root/AU/Documents/Treaties/text/Charter_on_statistics%20-%20EN.pdf

Council for European Social Science Data Archives (CESSDA) (2012) About CESSDA. Retrieved March 8, 2013 from the World Wide Web: http://www.cessda.org/about/members/

(CRDCN) The Canadian Research Data Centre Network (2011) Retrieved September 12, 2011 from the World Wide Web: http://www.rdc-cdr.ca/

DataFirst's website information (2011) Retrieved September 12, 2011 from the World Wide Web: http://www.datafirst.uct.ac.za/home/

International Household Survey Network (2011) Accelerated Data Program. Retrieved September 12, 2011 from the World Wide Web: http://www.ihsn.org/adp/

Kiregyera, B. (2005) A case and some actions for improving statistical advocacy in poor developing countries. *African Statistical Journal* 1, pp 70-84.

Lufumpa, C.L. & Mouyelo-Katoula, M. (2005) Strengthening statistical capacity in Africa under the framework of the International Comparisons Program for Africa (ICP-Africa). *African Statistical Journal* 1, pp 30-47.

NESSTAR website information (2011) Retrieved September 12, 2011 from the World Wide Web: http://www.nesstar.com/

Regional Reference Strategic Framework for Statistical Capacity Building in Africa (RRSF) (2006) *African Statistical Journal* 2, pp 131-134. Retrieved August 31, 2009 from the World Wide Web: http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/African.Statistical.Journal_Vol2_2.Articles_6.ReferenceRegionalStrategicFramework.pdf

South African Data Archive (SADA) (2011) Introduction. Retrieved September 12, 2011 from the World Wide Web: http://www.nrf.ac.za/sada/introduction.html

US Census Bureau. Census Bureau Information Quality Guidelines. Washington: United States Census Bureau. Retrieved March 7, 2013 from the World Wide Web: http://www.census.gov/quality/guidelines/index.html

Woolfrey, L. (2010) MPhil thesis, University of Cape Town (Unpublished). (Available from the author).

World Bank central microdata catalog. (2011) Retrieved September 12, 2011 from the World Wide Web: http://microdata.worldbank.org/index.php/catalog

(Article history: Available online 23 March 2013)