What does integration imply in choosing a unit of enumeration: enterprise, holding or individual? Does it matter? Perspectives from Africa.

Willis Oluoch-Kosura
University of Nairobi, and
CMAAE ,
Mebank Towers
Milimani Rd
Box 62882-00200
Nairobi, Kenya
Willis.kosura@aercafrica.org

Abstract

Statistical integration is advanced as a means of making data meet particular needs and ad value to the whole statistical data collection and management system. At the data collection level it can produce significant benefits by reducing the cost of statistical collection and the burden placed on respondents, whilst also increasing the value of outputs in terms of achieving consistency and accuracy. At the data processing level, integration enables the benefits of common technology, analytical methods, tools and processes to be fully exploited. Data from different sources and different times can be consolidated to allow for richer databases to be developed and meaningful comparative analysis and interpretation of results to be achieved. Integration implies that common statistical frames, definitions and classification are promoted and used in all statistical surveys to achieve harmony in enumeration units such as enterprise, holdings or individual. However, flexibility which may be necessary at times is lost. For instance agricultural data may require zonal based sample frames than household agricultural holdings. The periods for data collection may be agricultural seasons rather than calendar months. Thus integrating general household survey data with agricultural data may be complicated. This paper examines the issues of statistical integration from the African perspective and discusses the challenges, opportunities and the desirability of pursuing and achieving such integrated systems.

1. Introduction

Statistics play a major role in the scientific and development sectors of many nations. Within the development sector, reliable statistics are vital in describing the reality of people's lives and providing the evidence required to develop and monitor effective development policies; allows for the management of effective delivery of services by highlighting where resources are most needed; and provides the means of tracking progress and assessing the impact of different policies (Lipszyc, 2007). Moreover, up-to-date statistics leads to reliable information vital for the management of a region's natural resources and for dealing with regional development decisions that have a spatial context (Klosterman, 1995). While the role of statistics is recognized, the availability of quality statistics and their application in management is challenged by both technical and institutional issues facing many African economies. To remedy the situation, statistical integration is championed as a means of availing data due to its capability of bringing diversified data sources together for common good.

By definition, statistical integration implies a framework of applications, techniques and technologies (eFinity, 2006) for combining data from different sources in order to provide the user with a unified view of the data (Lenzerini, 2002) or the process of combining data from separate sources for making use of the information in estimating accurately the missing values in any of the single datasets (ARF, 2003) or an approach for enhancing the information content of separate statistical collections by ensuring consistency (Colledge, 1999). This is because it enables the replacement of disconnected sources of data or databases with an integrated structure that supports reliable and robust tools, processes and procedures for transforming disparate, unrefined data into understandable, credible, and easily accessible information for action (Polach and Rodgers, 2002). It success is attributed to the rapid adoption of databases with capacity to share or merge existing data repositories for the purpose of bridging the information gap (ARF, 2003).

While integration enhances data availability and compatibility, numerous challenges face its implementation in African countries. They range from how data are collected or situations surrounding the collection process, to institutional and technical factors. This paper explores the importance of statistical integration, challenges facing it and possible enhancing opportunities.

2. Importance of Statistical Integration

Statistical integration can provide several advantages. First, integration allows data produced from different sources, and at different times, to be brought together to provide richer datasets necessary for detailed and extensive analysis. Although this may have cost implications related to analysis, storage and dissemination, the availability of a large pool of data lowers the cost of planning and executing regular field data collection. This in turn lowers the burden placed upon respondents through constant interviews.

Secondly, owing to the need for consistency and quality, it promotes common standards in enumeration such as sampling designs, common definitions (standard concepts and variables e.g. what is a household), data classification (social, economic, geographic or environmental) and sample frames (e.g. household, an agricultural holding) (Kiregyera, 2001). As a result data

coherence is increased permitting for repeated collection of comparable data, in addition to providing opportunities for cross analysis, exchange and re-use of data. Standards, definitions and classifications also enhance data accuracy by promoting data scrutiny to ascertain their contents and quality. Poor quality data are either eliminated or transformed in the integration workflow through a process of data restructuring, cleansing, reconciliation and aggregation (White, 2005). Hence, data integrity and confidence in sharing among the different players is promoted. Moreover, consistent methodologies allow for comparative analysis between units of enumeration or same unit between time scales.

It also becomes possible to promote the use of common tools and processes in statistical analysis, methods of data storage (e.g. databases) that can be easily accessed across the board; as well as economical use of existing human capacity to prepare, analyse and interpret data for the common good. This in itself reduces the cost of specialized training needed in every department.

Integrated systems are also likely to provide a much detailed view of the situational analysis than using single data source (Hwang et al 2005). For example, integrating datasets within a geographic Information system (GIS) framework, it becomes possible to combine agricultural and socio-economic data to reveal disparities emanating from variations in household well-being (Rodgers, Emwanu & Robinson, 2006) or to reveal spatial-temporal patterns.

Nevertheless, flexibility is likely to be lowered by standards, definitions and classifications. For instance, incorporating the size of sampling unit (e.g. per hectare) would lower the flexibility of basing statistics on household agricultural holding thereby increasing time and efforts needed. Similarly, improving data accuracy by undertaking data collection during cropping seasons limits annual surveys based on interviewee recalling power or the possibility of conducting surveys during vacations when large numbers of educated person can serve as field personnel (Keita, 2004). Moreover, it makes it difficult to integrate household and agricultural data if the adopted standard sample frames do not correspond (Goguen, 2006). It is therefore important to pursue integration in a way that recognises such risks while maximising benefits.

3. Limitations to statistical integration

3.1 Limitations related to unit of enumeration

Some of the constraints to statistical integration emanate from the failure to decide on the right unit of enumeration for data collection, especially when considering panel or time series data needs. They include:

(i) Variation in data needs and sources: Normally, statistics are gathered through experimentation, surveys and censuses from households, individuals, or enterprises depending on the need in question. For example, an institution interested in agricultural performance would use censuses and surveys while one interested in administrative issues would use agricultural service data (FAO, 2005). The variation in need and sources of data make it difficult to decide on the best unit of enumeration. Moreover, because different institutions have varying problems and data needs, even under similar unit of enumeration (e.g. household) there are possibilities of variations in questions posed, type of questionnaires used and depth of coverage, as well as data collected that make data incompatible for integration.

(ii) Use of varying local units: African agriculture is characterized by existence of a large number of small subsistence farms with area cultivated depending on available manpower. Although agricultural holding has been recommended as the statistical unit (Kiregyera, 2001), surveys consider agricultural household (household in which at least one member operates an agricultural holding) rather than the holding itself, ignoring the aspect of size. This has led to lack of standards as varying units such as hectares, acres or other local units are employed. Similarly, data on production is estimated based on local units without standardization (Gutu, 2001) making it difficult to compare data from different households or regions.

(iii) Variation in production methods: The bulk of agricultural production comes from peasant farmer using a wide variety of farming techniques (Gutu, 2001) e.g. mono and mixed cropping and with varying spacing patterns driven by efforts to increase production and lower effects of natural calamities (bad weather, diseases and pests infestations). Hence, no two holdings may be similar to allow for comparisons and integration. As a result, sampling units based on factors other than ownership e.g. agricultural zonation would be devised for comparison purposes.

(iii) Surveys based on recalling power: Another drawback emanates from the appropriate timing to undertake data collection. Normally annual surveys are based on recalling power of interviewees despite significant cases of illiteracy (Keita, 2004). Time lapse between crops growth (seasons) and annual surveys (CBS, Kenya) has lead to inaccurate estimations and subjectivity on the part of respondent even when considering similar sampling frame, and this complicates statistical comparisons.

(iv) Definition of household: Although household has been preferred as unit of enumeration for agricultural census or surveys (Kiregyera, 2001), in most African economies it remains a major challenge because no clear definition of it has been provided. This is irrespective of the fact that a household in African context could be single, nuclear or extended and surveys using different definitions are likely to generate inconsistent data that are hard to integrate.

(v) Varying sampling methods: African institutions are faced with resource limitations, which make regular agricultural censuses impossible. Instead, sample enumeration is common (Keita, 2004). However, varying sampling designs or methodologies are applied in surveys depending on research goals and available resources, among random, stratified and multi-stage sampling. Varying methodologies leads to varying sampling units as well as obtained data even within same geographic region or population.

(vi) Diseases: There are numerous diseases including AIDS with devastating effects on African populations over short time periods. Such diseases have great impacts on agricultural activities and data collection especially when a household that forms a sampling unit for panel/time series study is wiped out such that no further data can be obtained. Change in sampling unit becomes inevitable otherwise data gaps unsuitable for integration occurs.

(vii) Migration and cultural barriers: Like agricultural systems, pastoralism is a major economic activity in many African countries and involves nomadism or migration at certain times of the year depending on resource availability. However, though migration is a major productive technique used to overcome numerous constraints posed by the environment (Tadingar, 1994), it

constantly disrupts data collection since target households cannot be traced. Despite nomadism, there is low levels of literacy and lack of formal education which creates communication barriers with the outside society. Even under settled conditions, some societies have customs that bar outside men from addressing women (Mogoa & Nyangito, 1999) such that when household heads (men) are not available, it becomes difficult to use such a households for data collection.

(viii) Insecurity and conflicts: insecurity situations like armed wars, banditry and tribal/clan conflicts experienced in several parts of Africa such as Darfur, Somalia, Chad or north-eastern Kenya, drives people out of their living areas, get killed or the situation rendered hostile to undertake data collection due to dangers posed to involved personnel. This implies targeted units of enumeration may sometimes not be reached calling for adjustments.

3.2 Limitations posed by institutional and technical factors

In addition to challenges emanating from variation in unit of enumeration at data collection phase, there are also technical and institutional issues that include:

(i) Policy and legal constraints: fears of integrated data systems result from legal constraints and copyright issues, lack of understanding as to whether integration will allow for access to datasets containing personal information that infringe on the Data Protection Act 1998, security reasons since some datasets are considered confidential and use is restricted to the concerned state/organization decision-making organ(s) or due to commercial value of data such that they can only be accessed at a fee (Jones & Taylor, 2004). Moreover, there exists varying national data policies with some countries availing their data freely while others hold high confidentiality or tight data access policies that discourage casual and unqualified data requests (FAO, 2003).

(ii) Weak national statistical systems: Many African countries have weak national statistical systems characterized by limited skilled manpower. This challenges integration through lack of effective capacity to generate new statistical data from existing ones or collect, analyze and manage data from different sources for use in decision support, especially when dealing with fast evolving scientific methods and technologies for data storage and management (FAO, 2003). Taking the case the Central Bureau of Statistics in Kenya (Table 1) (Central Bureau of Statistics, Kenya) we observe that most personnel hold lower qualifications, a situation that calls for upgrading to create capacity for effective statistical data management.

Table 1: Qualifications for Personnel at the Central Bureau of Statistics, Kenya

Post	No	Diploma/ Certificate	BA,BSc/ B.Phil	MA/MSc /M.Phil	Area of Specialization
Director	1			1	Economics
Deputy Directors	5	1	2	2	1 Demography; 2 Economics; 1 Statistics; 1 computing
Principal Statisticians	2		2		Statistics
Senior Stasticians	13	1	7	5	9 Statisticians; 2 Demography; 1 Economics; 1 Computing
Stasticians I.	42	2	37	3	35 Statisticians; 5 Computing; 1 Demography; 1 Geography
Snr Stat Officer	5	4	1		Statistics
Stat.Officer I	9	4	5		8 Statistics; 1 Cartography
Stat.Officer II	42	26	16		Statistics
Stat.Officer III	39	36	-?	-?	Statistics
Snr Stat. Assis I	12	10	-?	-?	Statistics
Cartographers	9	9			Cartography
Total	**179**	**93**	**70**	**11**	

(iii) Varying technological requirements for integration: The situation of limited skilled manpower is complicated by the existence of various data integration technologies with varying requirements e.g. each technology may have own user interface to the developer, own development environment, own metadata repository to document, and demanding own security and management frameworks, which makes it difficult to decide on the most appropriate technique to adopt (Meta Group, 2004) or the data integration technology specialists to engage.

(iv) Poor IT infrastructure: Many African countries have poor infrastructure for statistical integration owing to ineffective IT connectivity (White, 2005). This limits data sharing and communication of the results. Hence, countries are challenged in adopting the rapidly evolving data collection, processing and analytical techniques, storage and dissemination technologies especially in areas like GIS, satellite data processing and modelling (FAO, 2003).

(v) Data inconsistency and poor quality: Owing to the fact that most existing data were collected without common standards, definitions or classifications, there is high inconsistency among various data sources (FAO, 2003) while others are of poor quality (White, 2005) because of methods used in collection. Others are in poor formats and inadequate presentations that limit integration. For instance, most data lack metadata necessary in providing summary information such as sources and origin of data, methods of access and processing, fitness for use, adjustments made and their impacts on data integrity (Polach & Rodgers, 2006; Guptill 1999). Moreover, due to lack of standards for data collection, some data are not classified in to the various categories (numerical, qualitative, discrete, continuous, categorical (Hwang et al, 2005) or social, economic/geographic for faster integration.

(vi) Weak coordination and collaboration: Gutu (2001) notes that in many African countries, agencies collect, compile and disseminate statistical information with or without reference to the

National Statistical Offices. Agricultural censuses, surveys and other statistical inquiries are undertaken in isolation with no lack of understanding and co-ordination between statistical data producers. As a result, data inconsistencies among producers are common due to methodologies employed or units of enumeration used, which creates the need for enormous tasks in refining such data for integration purposes.

4. Opportunities for statistical integration in Africa

Despite the various challenges facing African countries in realizing statistical integration, there are opportunities that could be exploited in realizing the goal. First, with the growing IT development, Africa has witnessed tremendous introduction of new technologies such as the internet, CD-ROMS and databases that can allow for new ways of producing statistics by easing calculations and comparisons (Sverdrup, 2005). Moreover, the technologies can drastically reduce the costs of producing and disseminating statistics; allow for establishment of data dissemination centres or allow for wider and faster data dissemination than through printed publications. It would therefore be pragmatic to take advantage of such integrated systems enhancing technologies.

Furthermore, the growing internet and databases technology has created room for use of GIS in data integration especially when considering spatial information in decision-making (Jones and Taylor, 2004). Through its database framework, disparate data sources can be integrated and modelled into compatible ways for easier analysis, visualisation and sharing among users. For instance, GIS technologies have made it possible to integrate socio-economic and environmental data while still providing enough flexibility to derive information at desired geographic units and scales (Hung & Yasuoka, 2000). It is also possible to make conversions between areal units (e.g. agricultural holding to other spatial units) or create new areal units representing the intersection of the units through a variety of approaches. GIS has therefore made it practical to integrate agricultural data collected at households with other spatial related datasets for meaningful analysis.

Although there have been considerable investments in national statistical systems, there are new financing opportunities that countries can take advantage of to improve their statistical systems (Lipszyc, 2007). Such include the World Bank's Trust Fund for Statistical Capacity Building (TFSCB), which is vital in designing of National Strategies for Development of Statistics (NSDSs); STATCAP that can help in implementation; the African Development Bank and the UK's DFID assistance in designing or updating of NSDSs. Other international interventions include AFRISTAT that promotes regional and economic integration, by aiming for consistency and better comparability of statistical data; and the PARIS21 with its focus on assisting low-income countries implement National Strategies for the Development of Statistics (NSDSs) with a view to producing better national statistics by 2010. Such financing opportunities could be of help in improving the infrastructure of national statistical systems. Apart from support towards national systems, governments and/or agencies can learn about data collection, storage, integration and dissemination from the upcoming international frameworks such as the AfriSTAT of FAO.

There is increased data demand among research agencies and nations (FAO, 2003), which implies higher roles that integrated statistical systems would serve in availing quality data. Coupled with the increasing cooperation amongst nations or agencies currently being experienced, there is potential to pull resources together or generate common standards and procedures for enhancing integrated statistical systems. Moreover, the cooperation should provide platform for ensuring conflict resolution in order to open up conflict prone areas for data collection. There also exist large amounts of data collected by some research bodies or agencies in the region such as the CGIAR system, and some national and international research organizations. Dialogue for access should be initiated and the data used constructively as kick off point of the integration process since this would ease the task of obtaining past data.

Finally, having realized the challenges posed by subjective based surveys, research bodies should emphasize objective surveys by capturing unit sizes and production per unit. Moreover, it would be prudent to undertake a survey for developing dependable conversion factors for the various units used and come up with equivalent standard units. Through such it would be possible to standardise data collected using subjective methods and fill in the existing data gaps.

5. Conclusion

Data integration in African countries is a major challenge despite the achievements being attained globally. The challenge emanates from failure to have common unit of enumeration thereby leading to variation in data collected, which is driven by various considerations, to institutional and technical constraints. However, opportunities exist in solving the problem. What is most required for integration to effectively work is ensuring right unit of enumeration, standards, definitions, and classifications to guide the process of data collection, as well as devising prudent ways of overcoming the existing constraints, especially those related to technological development and associated capacity.

References

ARF, 2003, ARF guidelines for data integration, ARF, New York

CBS, undated, Status of food and agricultural statistics in Kenya, CBS, Nairobi.

Colledge, M.J.1999, **S**tatistical Integration through Metadata Management, *International Statistical Review / Revue Internationale de Statistique*, Vol. 67, No. 1, pp. 79-98

FAO, 2003, Agri-environmental information and decision support tools for sustainable development. Committee on Agriculture, 17th Session, FAO, Rome, Italy, www.fao.org/DOCREP/MEETING/005/Y8343e.html

FAO, 2005, Agricultural censuses and gender: Lessons learned in Africa, FAO Regional Office for Africa, Ghana.

Goguen, J. 2006, Information integration, databases and ontologies. www.cs.ucsd.edu/~goguen,

Guptill S.C., 1999, Metadata and data catalogues. In Longley P A, Goodchild M F, Rhind D W, and Maguire D J (eds) *Geographical Information Systems: Principles and Applications.* New York, John Wiley and Sons: 677–92

Gutu, S.Z, 2001, Developing agricultural statistics within the overall national statistical systems. Paper Presented at the Workshop on Strengthening Food and Agricultural Statistics in Africa, South Africa, 22-26 November 2001

Hung T. & Yasuoka, Y., 2000, Integration and application of socio-economic and environmental

data within GIS for development study in Thailand. AARS, GISdevelopment.net (www.gisdevelopment.net)
Hwang, D, Rust, A.G., Ramsey, S., Smith J.J, leslie, D.M., Weston A.D., Pedro A., Aitchison
Jones, M. & G. Taylor, 2004, Data integration issues for a farm decision support system, *Transactions in GIS,* 8(4): 459–477

Keita, N. (2004), Improving Cost-Effectiveness and Relevance of Agricultural Censuses in Africa: Linking Population and Agricultural Censuses. FAO Regional Office for Africa, Ghana, http://www.siea.sagarpa.gob.mx/mexsai/trabajos/t32.pdf
Klosterman R.E., 1995. The appropriateness of geographic information systems for regional planning in the developing world. Computer, Environment and Urban Systems, vol. 19, No. 1, pp. 1-13.
Lenzerini M., 2002, "Data Integration: A Theoretical Perspective". *PODS 2002*: 243-246
Meta Group, 2004, The future of data integration technologies. A Meta Group White Paper, www.metagroup.com/www.sunopsis.com
Mogoa, E.G.M. & M.M. Nyangito, 1999, Constraints to the delivery of animal health services in pastoral areas of Kenya:A review. The African Pastoral Forum, Working Paper No 20.
Polach, R.& M. Rodgers, 2006, The importance of data integration, IIM National, www.iim.org.au/national/htm/default.cfm
Rodgers D., T. Emwanu, & T. Robinson, 2006, Mapping poverty in Uganda using socio-economic, environmental and satellite data. http://www.fao.org/ag/pplpi.html
Sverdrup, U., 2005, Administering information: Eurostat and statistical integration. Working Paper No27, Centre for European Studies, University of Oslo, www. Arena.uio.no
Tadingar, T, 1994, Pastoral development in SSA: An integration of modern and indigenous technical knowledge. *The African Pastoral Forum*, Working Paper 2.
White, C., 2005, Data integration: Using ETL, EAI, And EII tools to create an integrated enterprise, BI Research, www.tdwi.org