Standards and tools for the dissemination of agriculture microdata: review and improvements

Matthew, Welch
The World Bank, Development Data Group
1818 H Street NW
Washington DC, 20433, USA
mwelch@worldbank.org

DOI: 10.1481/icasVII.2016.h46c

ABSTRACT

Large public and donor investment has led to an increase in the collection of microdata from surveys, census and administrative data. Preservation, discovery and access to these data are uneven and differ greatly by country and type of data. Extensive work by governments and the international community, through networks such as the International Household Survey Network (IHSN) and the Paris21 Accelerated Data Program (ADP), have provided the standards, guidelines and tools that have made it possible for many countries to preserve and disseminate microdata. One area, which is lagging, is in access to agriculture microdata. Agriculture microdata often possess both the characteristics of a household survey as well as an establishment survey, which can make the safe release of the data challenging. The paper discusses the tools developed by the International Household Network (IHSN) and the World Bank for microdata dissemination and includes recommendations on metadata standards as well as a discussion on guidelines and tools available for statistical disclosure control. Standards used for the documentation, preservation and dissemination of a typical household survey may not fit all types of data needed for the generation of agriculture statistics. The paper makes suggestions as to how the standards and tools could be improved to better suite agriculture surveys and in addition, non-survey sources of data for agriculture statistics.

Keywords: Metadata Standards, Disclosure Control, Dissemination, Tools

1. Introduction

There is a global movement towards more open access to data. Large public and donor investment has led to an increase in the collection of microdata from surveys, census and administrative sources. Many agencies provide access to aggregate, tabular or time series data. The preservation, discovery and access to the microdata underlying these sources are uneven and differ greatly by country and type of data. The release of the underlying microdata is important as it allows researchers and policymakers to replicate officially published results, provide feedback for improvement of future surveys, generate new insights into issues, avoid duplication of surveys and provide greater returns to the investment in the survey process.

Extensive work by governments and the international community, through networks such as the International Household Survey Network (IHSN), coordinated by the World Bank, and the Partnership in Statistics for Development of Statistics in the 21st Century (Paris21) have provided standards, guidelines and tools that have made it possible for many countries to preserve and disseminate microdata. These programs have resulted in the documentation of thousands of surveys by countries and agencies across the world. While these programs have ensured substantial improvements in the preservation of data and dissemination of good quality metadata, many agencies are still reluctant to allow access to the microdata. This is particularly true of agriculture microdata. The Global Strategy to improve Agricultural and Rural Statistics (2010) identifies six common areas where agriculture statistics in developing countries face problems. The last two are of particular interest in the context of this paper.

- Limited staff and capacity of the units that are responsible for collection, compilation, analysis, and dissemination of agricultural statistics.
- Lack of adequate technical tools, statistical methodology, and survey framework to support data-production efforts.
- Insufficient funding allocated for agricultural statistics from development partners and national budgets.
- Lack of institutional coordination, which results in the lack of harmonized and integrated data sources.
- Lack of capacity to analyze data in a policy perspective, which results in a significant waste of resources, as large amounts of raw data are not properly used.
- Difficulty for data users in accessing existing data with no metadata or indication of quality.

The Global Strategy acknowledges that building and maintaining technical capacity is a difficult task and requires a long-term plan. One of the recommendations includes that capacity building should include support to the dissemination of results. This could be extended to, or more specifically state, that capacity building to support microdata dissemination should be built. One of the key benefits of dissemination of the microdata is that it widens the pool of experts immediately who can analyze the data to gain the best and timeliest insights possible. It is not possible for the staff of statistical offices to respond to all analysis requests. Releasing the microdata frees up capacity for core activities by allowing users of the data to generate their own analysis.

In order to ensure that the released microdata are properly used they should be released with detailed metadata, which describes the full survey process, the data and any assessments of quality.

The following sections discuss the tools developed by the World Bank\IHSN to help countries and agencies disseminate microdata and discusses, and advocates for, the use of an international

standard for the documentation and creation of metadata for survey data called the Data Documentation Initiative (DDI).

2. A Standard for Microdata Documentation

There are a number of reasons commonly cited for why the creation and agreement on standards has benefits. These include allowing the interoperability of systems, lowering the cost for developing tools, improving quality. The most commonly used standard for the description and documentation of survey data is the Data Documentation Initiative (DDI). The DDI had its origins some 20 years ago at the University of Michigan and is now an international standard developed by a large community of practice. The standard is used by most social science archives around the world as well as by a large number of international and country agencies that collect and disseminate survey data.

The DDI standard provides a structured checklist of what needs to be documented to fully understand the entire survey lifecycle, the microdata and its quality.[1] The comprehensive and structured nature of a survey documented according to the DDI standard helps agencies preserve data and makes it easier for users to understand what the data are measuring and how the data were created.

The DDI is comprised of two standards. The DDI Lifecyle and the DDI Codebook.[2] The Lifecycle accommodates the advanced documentation of the survey lifecycle from the conceptualisation of the survey through to analysis. The complexity of the DDI Lifecyle has meant that its uptake has been limited to only a few advanced archives and agencies. This led the World Bank\IHSN to champion an effort to simplify the standard. The DDI Codebook is based on feedback coming mostly from developing countries and as a result, the DDI Codebook is a more bottom-up standard.

The relative ease with which the DDI Codebook can be understood made it possible for the World Bank and IHSN to easily develop tools and guidelines for preservation and documentation of surveys. As a result, the standard is now being implemented in 130 national agencies in 88 countries, and 18 international or other organizations. National Statistics Offices concerned about preservation of their data and institutional memory combined with the World Bank\IHSN have documented over 5000 surveys using the DDI standard.

The standard is suitable for the documentation of any type of survey or census data and as such is suitable, and has been widely used, for the documentation of agriculture surveys and agriculture census data.

Looking forward, the core elements of the DDI Codebook will likely remain the standard of choice for the documentation of surveys from most of the agencies currently using it. Given though that the standard was designed for survey and census data and given that many agricultural statistics are now being generated from other sources of data or are merged with other types of data it will be necessary to adapt or add to the current standard in order to accommodate these sources. Examples

[1] The Quick Reference Guide for Data Archivists produced by the IHSN gives a description of key fields in the DDI and how the World Bank\IHSN have implemented it.
[2] The DDI Alliance: http://www.ddialliance.org/

of other sources of agricultural statistics are administrative data, remotely sensed data, satellite imagery, cadastral surveys, and time series data.

The DDI Alliance are currently working towards a new version of the DDI (DDI 4). The World Bank is working with the DDI alliance to ensure the standard meets the changing needs of users in developing counties, while still maintaining its relative simplicity for implementation. DDI 4 will incorporate all the elements from DDI Codebook and DDI Lifecycle in a simpler view. It will make it easier to use DDI for geospatial, administrative and time series data. This will be a major benefit to much data used for agriculture statistics. DDI 4 will come with a feature called profiles which will allow the user to pick a subset of elements for specific types of data. For example, for geospatial data, a DDI template (schema) could be created only for geospatial data or a combination of schemas could be created for multiple data types.

3. Tools for Documentation and Dissemination

The previous section discussed the needs for, benefits of, and the successes, which the World Bank and IHSN have had in rolling out standards for the documentation of microdata. A large part of the success in the adoption of the DDI by so many agencies has been the availability of easy to use free tools for the documentation and dissemination of surveys using the DDI standard.

The typical DDI document consists of an XML schema. While it is possible to write a DDI document directly in XML this is not easy nor convenient. To make the creation of DDI documents quick and easy requires software tools that allow the creation of DDI documentation without the user having any knowledge of XML. DDI documentation can be transformed into multiple human readable formats such as PDF reports and web based pages. The DDI lends itself most powerfully to the creation of online searchable catalogues. The World Bank and IHSN have been using and supporting a freeware DDI compliant editor as well as an open source online DDI compliant survey cataloguing and dissemination platform since 2006.

3.1 A DDI editor

The World Bank and IHSN contributed to the development of a DDI compliant metadata editor (The Nesstar Publisher) produced by the Norwegian Social Science Data Archive (NSD). This is Windows based software, which supports multiple languages. It is offered as Freeware by the NSD and has been the key vehicle through which the World Bank and IHSN have implemented the DDI standard since 2006. The availability of this easy to use free editor has led to the preservation and documentation of thousands of surveys, which might otherwise have been hidden or lost.

While the Nesstar Editor will continue to remain available for the near future, it is not anticipated that the NSD will develop the Editor any further. Given the earlier discussion on the need for further development of the DDI standard to incorporate other types of data there is a need to plan for and develop a new editor that can accommodate the new standard and new sources of data. While some advanced data centres have developed their own editors and some commercial editors are now available; there remains a need for the availability of a free, commercial-quality DDI compliant editor. To solve this problem The World Bank have begun the process of developing a new DDI editor. The new editor will support future versions of DDI and will have the added ability to support other standards such as Geospatial ISO 19139 and Dublin Core as well as custom

designed schemas. The new editor will run on multi-platforms and will be distributed as open source software. The flexibility of the new editor combined with World Bank and IHSN support to countries is expected to broaden the ability for new types of data (relevant to the generation of data such as those used for agricultural statistics) to be documented, described, preserved and ultimately disseminated.

3.2 A Dissemination Platform

Data have their broadest use and utility if they can be discovered and used by a broad range of users. The most cost effective and efficient way to do that is to display the information online in such a way that makes it possible for users to discover, browse, search and download the metadata and microdata from a single online platform. Some advanced agencies have built their own custom microdata and metadata catalogues, and there are commercial products designed specifically for survey microdata available.[3] The success, which the World Bank and IHSN have had in helping countries, build capacity for the dissemination of data and DDI compliant metadata has however come through making available free (open source) catalogue software developed by the World Bank and supported by the IHSN partners. The software is called the National Data Archive (NADA). It is a multi-platform DDI compliant open source PHP application. The NADA allows users to browse, search, filter and download metadata and microdata. It also comprises a full featured secure administrative interface that allows managers of the system to apply access control polices at the survey level and to monitor and administer requests made by users. The full list of countries and organizations that the World Bank and IHSN have supported in implementing the NADA platform can be found in Appendix 1.

The NADA platform was originally designed to support only the DDI standard and survey and census microdata. User feedback and requests from users have always guided the development path of the NADA platform. The needs of users have changed over time as statistics are generated from new sources or combinations of sources of data. This is especially true of agricultural statistics where data come from many sources and where new data sources are augmenting or replacing survey and census data as a source. To address changing needs the World Bank have begun the development of new version of the NADA. The new version will accept and facilitate, in addition to new and existing versions of DDI, the display and dissemination of multiple metadata standards and data types. Of importance to agricultural statistics, this includes administrative data, time series data and geospatial data.

3.3. Tools for privacy protection

The release of microdata is important, as it allows researchers and policymakers to replicate officially published results, generate new insights into issues, avoid duplication of surveys and provide greater returns to the investment in the survey process.
The release of microdata poses privacy challenges to the producer. Agriculture microdata in particular often possess both the characteristics of a household survey as well as an establishment survey, which can make the safe release of the data more challenging.

The dissemination of agriculture microdata is poor in developing countries. One of the key reasons why these data are not widely distributed relates to the difficulty many countries face in applying statistical disclosure control (SDC) or privacy protection measures to the data. See the review by the Global Strategy to improve agricultural and rural statistics (2014).

[3] Two of the best known commercial products are the Nesstar Server and products from Colectica

While the release of microdata poses often-difficult disclosure control problems this is not the case for all types of data related to agriculture statistics. There are many examples where data from agriculture surveys have been released. One of the prime examples is the LSMS-ISA data produced by the World Bank. Also, a review of data catalogs in many developing countries shows that developing countries are successfully disseminating agriculture related data[4]. Where the data do present challenges for privacy there may be technical solutions and tools that could overcome some of these hurdles and allow the release of traditionally more sensitive microdata.

These strategies usually involve a combination of enabling legislation, the application of appropriate access terms and statistical disclosure control (SDC) methods.[5]

The World Bank and IHSN programs have provided considerable guidance to countries on the creation enabling legislation and dissemination policies for microdata dissemination. This has addressed part of the problem. Countries consistently list the lack of capacity and knowledge of SDC methods for privacy control as a remaining barrier to greater release of microdata. This is particularly the case for many types of agriculture data.

To help address this problem the IHSN, PARIS21 (OECD), Statistics Austria and the Vienna University of Technology and the World Bank has contributed to the development of an open source software package for SDC, called sdcMicro.

The package was developed for use with the open source R statistical software, available from the Comprehensive R Archive Network (CRAN) at http://cran.us.r-project.org. The package includes a comprehensive suite of methods for the assessment and reduction of disclosure risk in microdata. For users who are familiar with R, sdcMicro offers a very powerful and free tool to treat microdata for safe release. For those users who are not familiar with R the World Bank and the IHSN are developing a graphic user interface for sdcMicro that will remove much of the need to know R and free users to simply apply the methods as required. This should be available towards the end of 2016.

The proper implementation of SDC methods requires expertise and experience in Statistics and in the area of SDC. Many advanced agencies have expertise in this specialized area, but this is much less the case in developing agencies. The provision of the free tools will go a long way to improving the ability of agencies to apply SDC methods, but without guidance and capacity building uptake may still be low. To this end, the World Bank and IHSN have produced a number of guides to the SDC process and the practical application of the sdcMicro package. [6]

It should be stressed that SDC is only one part of the data release process, and its application must be considered within the complete data release framework. The level and methods of SDC depend on the laws of the country, the sensitivity of the data and the access policy (i.e., who will gain access) considered for release. The provision of the free SDC package and guidelines create an enabling environment for unlocking more microdata including agriculture microdata.

[4] See appendix 1
[5] For an in-depth summary of access polices, and dissemination of microdata files see The IHSN working paper written by Dupriez and Boyko (2010)
[6] See: http://www.ihsn.org/home/software/disclosure-control-toolbox

4. Conclusion

The success which the World Bank and its IHSN partners have had in assisting countries and agencies build capacity for the documentation, preservation and dissemination of microdata has, in large part, been due to a strategy of building, disseminating and supporting free tools. The thread linking the success of the tools has been the existence of an easy to understand and implement international standard – DDI.

These tools, and the support for them, have improved the accessibility to good metadata and microdata for thousands of surveys. There are still countries that despite having the necessary platforms and tools for dissemination do not release microdata for all or certain types of surveys. Agriculture microdata are one of the data types that are not widely disseminated and where solutions need to be found in order to maximize the investment in the collection of these data and for improvement in surveys and policy. A commonly mentioned reason for not distributing microdata from agriculture surveys has been privacy protection. To help solve this problem the World Bank and the IHSN have supported the development of a free tool for the implementation of SDC methods and developed guidelines to support this. This lowers the barriers to release and opens the door for more countries to release microdata.

With increasing amounts of data for agricultural statistics coming from sources other than surveys and census there is a need to adapt current tools and standards to incorporate these new sources. The World Bank and IHSN are constantly improving the available tools to ensure they accommodate the changing needs of producers and users of data. The World Bank is also actively working with the DDI alliance to ensure that documentation standards remain accessible and relevant.

REFERENCES

Benschop T., Machingauta C., Welch M. (2015). Statistical Disclosure Control for Microdata: A Practice Guide. IHSN. http://ihsn.org/home/sites/default/files/resources/Statistical%20Disclosure%20Control%20for%20Microdata_0.pdf

Dupriez O., Boyko E. (2010). Dissemination of Microdata Files Principles, Procedures and Practices. IHSN Working Paper, No 005. http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP005.pdf

Dupriez O., Greenwell G. (2007). Quick Reference Guide for Data Archivists. IHSN, http://www.ihsn.org/home/node/544

Global Strategy to improve agricultural and rural statistics, (2010). Providing Access to Agriculture Microdata: A Guide. http://gsars.org/wp-content/uploads/2014/09/Providing-Access-to-Agricultural-Microdata-Guide.pdf

World Bank, Food and Agriculture Organization of the United Nations and United Nations (2010). Global strategy to improve agricultural and rural statistics. Report No. 56719-GLB. Washington, D.C.: World Bank.

Appendix 1: NADA Catalogues around the World. Distribution of agriculture data through NADA catalogues

NADA Catalog	Agriculture		NADA Catalog	Agriculture	
	Metadata Available only	Meta & Microdata Available		Metadata Available only	Meta & Microdata Available
Angola / Instituto Nacional de Estatística (INE)			Liberia / Liberia Institute of Statistics & Geo-Information Services (LISGIS)		
Benin / INSEA			Malawi / National Statistics Office (NSO)		X
Bhutan / National Bureau of Statistics (NBS)			Mali / Institut National de la Statistique (INSTAT)	X	
Bolivia / Instituto Nacional de Estadística (INE)	X		Mauritania / Office National de la Statistique (ONS)		
Botswana / Statistics Botswana			Mauritius / Central Statistics Office (CSO)		
Burkina Faso / Institut National de la Statistique et de la Démographie (INSD)		X	Mexico / Instituto Nacional de Estadística y Geografía (INEGI)		
Burundi / Institut de Statistiques et d'Etudes Economiques du Burundi		X	Mongolia / National Statistical Office (NSO)		
Cambodia / Cambodia National Institute of Statistics (NIS)			Mozambique / Instituto Nacional de Estatistica (INE)	X	
Cameroon / Institut National de la Statistique (INS)		X	Namibia / Namibia Statistics Agency (NSA)		
Cape Verde / Instituto Nacional de Estatística (INE)			Nepal / Central Bureau of Statistics (CBS)	X	
Colombia / Departamento Administrativo Nacional de Estadistica (DANE)	X		Niger / Institut National de la Statistique (INS)		X
Costa Rica / Instituto Nacional de Estadística y Censos (INEC)			Nigeria / National Bureau of Statistics (NBS)		X
Cote d'Ivoire (Ivory Coast) / Institut National de la Statistique (INS)			Peru / Instituto Nacional de Estadística e Informática (INEI)	X	
Dominican Republic / Oficina Nacional de Estadística (ONE)			Philippines / National Statistics Office (NSO)		X
Ecuador / Instituto Nacional de Estadistica y Censos (INEC)		X	Rwanda / National Institute of Statistics Rwanda (NISR)		X
Equatorial Guinea / Instituto Nacional de Estadística (INEG)			Saint Lucia / Central Statistics Office		
Ethiopia / Central Statistical Agency (CSA)		X	Senegal / Agence Nationale de la Statistique et de la Démographie (ANSD)		
Gambia / Gambia Bureau of Statistics (GBoS)	X		Sierra Leone / Statistics Sierra Leone (SSL)		
Ghana / University of Cape Coast (UCC)			Somalia / Puntland-Ministry of Planning and International Cooperation		
Ghana / Ghana Statistical Service (GSS)		X	Somalia / Somalia-Ministry of Planning and International Cooperation		
World Food Programme (WFP)			South Africa / University of Cape Town / DataFirst		
World Bank - Microdata Library			Sri Lanka / Department of Census and Statistics (DCS)	X	
International Household Survey Network (IHSN)			Sudan / Central Bureau of Statistics (Arabic Language)		
Secretariat of the Pacific Community (SPC)			Sudan / Central Bureau of Statistics (English Language)		
Guinea / Institut National de la Statistique (INS)		X	Togo / Direction Général de la Statistique et la Comptabilité Nationale		X
Guinea-Bissau / Instituto Nacional de Estatística (INE)			Tunisia / Institut National de la Statistique (INS)		
Honduras / Instituto Nacional de Estadistica (INE)	X		Uganda / Uganda Bureau of Statistics (UBoS)		X
India / Ministry of Statistics and Programme Implementation (MOSPI)		X	United Republic of Tanzania / National Bureau of Statistics (NBS)		X
Indonesia / Badan Pusat Statistik (BPS)		X	Uruguay / Instituto Nacional de Estadística (INE)		
Jordan / Department of Statistics (English Language)			Vanuatu / Vanuatu National Statistics Office (VNSO)	X	
Jordan / Department of Statistics (Arabic Language)			Viet Nam / General Statistics Office (GSO)	X	
Kenya / Kenya National Bureau of Statistics (KNBS)			West Bank and Gaza / Palestinian Central Bureau of Statistics	X	
Laos / Lao Statistics Bureau (LSB)		X	Zambia / Central Statistical Office (CSO)		X