3 March 2014

Microdata dissemination best practices[1]

BACKGROUND

At its 16th session (Vienna, 2010), the CCSA endorsed a proposal presented by UNODC to review the practices of microdata release in international organizations. A Task Team was established with the aim to guide a debate in the CCSA on how best international organizations can manage the access to microdata while protecting confidentiality, and to compile best practices in developing policies and methodology on the access to micro-data in international organizations.

A desk review of available tools, guidelines and microdata dissemination practices by CCSA agencies was conducted. Only a few of the CCSA agencies publish microdata. UNICEF, WHO and the World Bank disseminate microdata on line. Eurostat does not publish microdata, but provides access to microdata for research/scientific purposes on the basis of legal agreements signed with members of the European Statistical System. Other CCSA agencies publish metadata related to survey or census operations (ILO, UNFPA, UNSD, others).

All CCSA member agencies can or do play an important role as providers of recommendations and financial or technical support to microdata dissemination by national agencies.

THE DEMAND

The demand for microdata is growing and becoming more diverse. This diversity is visible in a broadening of the audience but also in the way in which the microdata are being used. Increasingly, as more Open Data information becomes available, users are seeking to combine data from diverse sources for better measurement and impact. Limited quantifiable measures such as web usage statistics obtained from microdata repositories provide anecdotal evidence that this demand is significant. While still predominantly originating from rich countries and international organizations, it appears to be increasing in developing countries. A more systematic and comprehensive assessment of the demand (i.e., by category of users, country of origin, datasets of interest, purpose of use, level of satisfaction, etc) would allow data producers to improve their data dissemination services.

THE SUPPLY

The benefits of microdata dissemination are well known and broadly accepted. Sharing microdata fosters diversity of research, increases transparency and accountability, and can mitigate duplication of data collection work and increase the quality of data through feedback received from data users. But the

[1] This note was initially drafted by the World Bank and presented at the 22nd CCSA session (Ankara, 2013) then improved based on input received from Eurostat, FAO, UNECE, UNICEF.

associated costs and risks are not negligible, the main one being the reputational risk in case of violation of privacy protection rules and regulations. While there has been no major incidents reported so far, this risk cannot be ignored. In response to these risks many data producers and depositors have adopted a conservative approach by severely limiting or excluding access to their microdata.

The Open Data and Open Government initiatives, combined with the positive experience of organizations that do disseminate microdata have however created a momentum for more openness. Adopting international standards and good practices of microdata dissemination offers greater guarantees that the various technical, ethical and legal issues will be properly addressed, and in so doing mitigate risk.

Key principles and organizational best practices

Many national and other agencies refer to the United Nations Fundamental Principles of Official Statistics when setting up their data dissemination policies. Best practices should be compatible with these principles. The sixth principle governing International Statistical Activities states that "Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes." The strict confidentiality is often invoked as a reason not to share any microdata. In 2006, a task force set up by the Conference of European Statisticians assessed the implications of the principle and produced *Guidelines and Core Principles of Confidentiality and Microdata Access* (UNECE, 2007) in which they suggested that:

- "It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected. (...) Making available microdata for research is not in contradiction with the sixth UN Fundamental Principle as long as it is not possible to identify data referring to an individual."
- Microdata should only be made available for statistical purposes: The aim must be to derive statistics that refer to a group (of persons or legal entities), not to specific individuals.
- Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.
- The procedures for researcher access to microdata should be transparent, and publicly available. This is important "to increase public confidence that microdata are being used appropriately and to show that decisions about microdata release are taken on an objective basis."

The Organisation for Economic Co-operation and Development (OECD) also defined a set of core principles related to microdata dissemination in their *Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007). The list of principles below is adapted from these guidelines:

- ***Openness:*** Openness should not be understood as "unrestricted access" or "open data"; it means that access must be provided on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination.

Note: All CCSA member agencies who publish microdata provide them free of charge.

- ***Transparency:*** Detailed metadata must be provided, and the specifications of conditions attached to the use of the data should be internationally available in a transparent way, ideally through the Internet.

  Note: All CCSA member agencies who publish microdata provide them on-line with rich metadata. The access conditions attached to the data vary but are clearly specified. None of the organizations provide microdata without any restrictions. Eurostat produces "Scientific Use Files" which are shared for scientific purposes only with researchers affiliated to organization recognized by Eurostat as research entities (universities, research institutions or research departments in public administrations, banks, statistical institutes, etc). UNICEF, WHO and the World Bank disseminate microdata as Public Use Files or Licensed Files, accessible to registered users. In all cases, the terms of use associated with microdata include obligations and restrictions (e.g., prohibiting attempts to re-identify respondents or selling or transferring the data to others). A compilation of practices for the formulation of access policies and terms of use is available in the International Household Survey Network Working Paper No 5 (Dupriez and Boyko, 2010).

- ***Legal conformity and protection of privacy:*** National laws and international agreements, as they pertain to the protection of privacy, directly affect data access and sharing practices. These must be taken into account in the formulation of data access arrangements. Microdata are typically obtained from households, individuals, firms or facilities against a commitment to keep the data confidential. Unless formal consent has been provided by the respondent (which is rarely the case) data can only be disseminated after being properly treated to ensure the risk of disclosure is minimal, i.e. the data are anonymized.

  Note: Various methods exist to quantify disclosure risk in microdata, but all of them rely on arbitrarily-defined scenarios. There is no universally-accepted method for measuring the risk, and no recommended thresholds for defining what an acceptable disclosure risk level would be. Decisions to declare microdata "fit for dissemination" are generally based on contextual factors (sensitivity of the data, legislation, reputational risk, potential consequence for the re-identified respondents, political context, etc).

- ***Protection of intellectual property:*** Data access arrangements must consider the applicability of copyright or of other intellectual property laws that may be relevant.

  Note: Microdata obtained (and in some cases published) by international organizations are often produced through data collection activities implemented and funded by multiple national and international partners. The ownership of the resulting data is not always clearly identified. The rights or obligations to disseminate microdata should be

explicitly defined in funding agreements and contracts. Good practice and models are available.[2]

- ***Interoperability:*** Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. Access arrangements, should pay due attention to the relevant international data documentation standards.

  Note: The DDI metadata standard is the standard adopted by most specialized microdata libraries and many statistical agencies. It was also adopted by several international organizations. See more about the DDI standard below.

- ***Quality:*** The value and utility of data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organizations, should pay particular attention to ensuring compliance with explicit quality standards.
- ***Security:*** Specific attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of data.
- ***Accountability:*** The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and funding agencies.

Technical standards and best practice

Best practice and standards in microdata curation —documentation, cataloging, anonymization and dissemination, and preservation—mainly come from the academic world.

- **Documentation**. Compliance with international metadata standards is crucial for ensuring exchangeability of metadata and for promoting collaboration or coordination in the development of microdata curation tools. The standard most commonly used for the documentation of microdata is the Data Documentation Initiative (DDI) by the DDI Alliance. The standard is used by most of the large social science data archives around the world, by many national statistical agencies, and by international organizations (including FAO, ILO, UNICEF, WFP, WHO, and the World Bank). A number of free tools exist for creating DDI metadata, such as a free DDI editor developed by NESSTAR Ltd and supported by the IHSN. The DDI standard complements the SDMX standard. The (long) process of gaining ISO certification for the standard has recently been initiated. In the meantime, adopting the DDI as a UN recommended standard for microdata documentation would bring the UN institutions in line with current best practice.

  Ideally, the DDI standard should be complemented by the adoption of a common, multilingual taxonomy of topics. Such a commonly accepted taxonomy does not exist yet, but could be created based on taxonomies developed by academic data centers (such as the Council of

[2] See for example the standard memorandum of understanding developed by UNICEF for the Multiple Indicator Cluster Surveys (available in five languages at http://www.childinfo.org/mics5_planning.html)

European Social Science Data Archives - CESSDA) or by national agencies (Statistics Canada's taxonomy provides a particularly relevant model).

- **Cataloguing**. Publishing detailed metadata in on-line searchable catalogs is important to make data discoverable. Compliance with the DDI standard makes it considerably easier. Open source DDI compliant cataloging applications already exist such as the open source **Na**tional **D**ata **A**rchive (NADA) developed by the IHSN or DataVerse developed by Harvard University.
- **Anonymization.** This is an area where academically well-grounded methods exist. Tools exist for the measurement and reduction of disclosure risk, such as the open source, R-based sdcMicro application (Templ et al, 2012) or µArgus (a freeware, soon to be open source). But no international standards exist for their implementation. Institutional and national practices are typically not documented or shared. The reasons for this are that the methods used for anonymizing microdata are very contextual to the type of data being anonymized, and that disclosing detailed information on the methods may provide useful information to those trying to defeat the protections. Eurostat publishes broad information on their anonymization methods on their website. Such transparency is good practice as it makes researchers aware of the content and limitations of the microdata. Useful but somewhat outdated information can also be found in the *Report on Statistical Disclosure Limitation Methodology* by the US Federal Committee on Statistical Methodology (2005). An international review of disclosure risk management practices by statistical agencies, research centers and other data repositories would be tremendously useful.

  The IHSN and the World Bank have initiated a research project aimed at evaluating, using a large collection of household surveys, the impact of various anonymization algorithms on the risk level and resulting information loss. This work should result in the production of a "practice manual".
- **Dissemination**. Cognizance needs to be given to the fact that not all data are the same. Dissemination policies need to be developed that are clear but also flexible enough to cover the full range of issues such as data ownership, legal and ethical responsibilities and sensitivity of data. Some data are more sensitive than others, and as such dissemination policies must provide for multiple access policies to accommodate various types of datasets. Typically, five levels of accessibility are considered: open access (no restriction), direct access or Public Use Files (some restrictions on use, but no screening of users), Research Use Files (or Scientific Use Files, or Licensed Files), availability only in an enclave, and no access authorized.
- **Preservation**. Organizations which disseminate microdata and the related metadata are also often responsible for their long term preservation. Preserving digital content is not a trivial exercise. Procedures and infrastructures must be put in place to protect data against hardware and software obsolescence (regular migration of datasets to new media and formats), system failures, human errors and other hazards. The IHSN Working Paper on *Principles and Good Practice for Preserving Data* by the Interuniversity Consortium for Political and Social Research (2009), provides guidelines and multiple references to recommended practices and standards such as the Open Archival Information System (OAIS).

Note: The preservation practices by international organizations are not documented; it is unlikely that they comply with best practices.

Implementing technical best practices has a significant financial cost. Maintaining a best-practice dissemination system requires continued and dedicated budgetary planning and appropriate funding. The cost can be significantly reduced by adopting common tools and standards, which opens the door to options such as common data repositories.

POSSIBLE AREAS OF COORDINATION

The IHSN has invested resources and has established multiple partnerships to develop and promote international technical best practices of microdata management and dissemination. This includes participation in the DDI Alliance to further develop the DDI standard, the development of free/open source software for the documentation, cataloging, anonymization and dissemination of microdata, and the maintenance of a central metadata catalog. The IHSN is well placed to keep playing a lead role in the maintenance and regular improvement of these technical tools and guidelines. This includes coordinating discussion and providing guidelines to agencies and NSO's in new areas where demand is increasing, and where the creation of new and better products can lead to improved impact on Open Government and research. One such area might be in coordinating discussions on guidelines for agencies and NSO's on preparing and providing microdata in combination with other data sources in ways that maintain legal conformity and address privacy. The IHSN may also be well positioned to conduct a compilation of national practices of microdata dissemination around the world, with particular focus on the legal obstacles to microdata sharing. The IHSN is open to international and other organizations to join in these efforts.

Further, an OECD Expert Group on international collaboration in microdata access is finalizing recommendations related to access to microdata across borders.

The CCSA is invited to recommend the adoption of the DDI metadata standard for microdata by International Agencies and to promote its use by member countries.

References

- Inter-university Consortium for Political and Social Research (ICPSR). 2009. “Principles and Good Practice for Preserving Data”, International Household Survey Network, IHSN Working Paper No 003. December 2009.
  http://www.ihsn.org/home/node/121
- Olivier Dupriez and Ernie Boyko. 2010. Dissemination of Microdata Files - Principles, Procedures and Practices. IHSN Working Paper No 005
  http://www.ihsn.org/home/node/120
- Organisation for Economic Cooperation and Development (OECD). 2007. “OECD Principles and Guidelines for Access to Research Data from Public Funding”.
  www.oecd.org/dataoecd/9/61/38500813.pdf
- UK Data Archive, University of Essex. 2009. “Managing and Sharing Data. A Best Practice Guide to Researchers”, second edition.
  www.dataarchive.ac.uk/news/publications/managingsharing.pdf
- United Nations Economic Commission for Europe (UNECE) and Conference of European Statisticians. 2007. “Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice”.
  www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf
- US Federal Committee on Statistical Methodology. 2005. “Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology”.
  www.fcsm.gov/working-papers/spwp22.html

Links

Microdata in the CCSA member agencies’ websites

- EUROSTAT: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/introduction
- UNICEF: www.childinfo.org
- World Bank: http://microdata.worldbank.org
- World Health Organization: http://apps.who.int/healthinfo/systems/surveydata/index.php/home

International Household Survey Network (IHSN): www.ihsn.org

DDI Alliance : www.ddialliance.org