WP.41
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (viii): International projects, groups and forum dealing with data access, release and related methodologies

Metadata standards to support controlled access to microdata

Prepared by Wendy Thomas (University of Minnesota), Arofan Gregory (Open Data Foundation), U.S.A. and Alistair Hamilton (Australian Bureau of Statistics), Australia

Metadata standards to support controlled access to microdata

Wendy Thomas,[*] Arofan Gregory, [**] and Alistair Hamilton [***]

[*] MPC, University of Minnesota, Minneapolis, MN 55455, USA, wlt@umn.edu

[**] Open Data Foundation, Tucson, AZ 85704, USA, agregory@opendatafoundation.org

[***] Australian Bureau of Statistics, Belconnen, ACT 2617, Australia, alistair.hamilton@abs.gov.au

1 Considerations in providing access to microdata

In an increasingly complex world, requests for wider access to the microdata collected by National Statistical Offices (NSOs) continue to rise. Aggregated data provide a limited picture and do not support the range of statistical analyses that can be accomplished through use of the microdata. Requests are coming from members of the social science research community as well as from government analysts seeking to undertake multifaceted (e.g., triple bottom line) evidence-based policy development and evaluation. It appears to be the case that, in principle, NSOs welcome, and would like to support, such requests.

A typical core role for an NSO is to support informed decision-making, research, and discussion within governments and the community. Providing appropriately managed access to microdata for research purposes can represent richer fulfilment of this role. In some cases allowing researchers to "self serve" from microdata, under controlled conditions, by selecting and analysing the data specifically relevant to their needs represents a more cost-effective strategy than requiring NSOs to produce large, pre-confidentialised datasets containing low-level aggregates just in case some of those aggregates may prove useful for an unknown external researcher in future.

Survey data are expensive resources to create, both from the perspective of the NSO and in terms of what is required from the load on individual respondents/providers. Appropriately managed, provision of access to microdata for research purposes may substantially increase the public benefit realised from this investment. In some cases it may prevent, or reduce, the need for the researcher to survey individuals or businesses separately. Transnational microdata access has the potential to support better informed international comparisons and benchmarking. The ability for one NSO to discover and explore the structure of the microodata collected by its counterparts around the world can assist in survey design, including designing for consistency and comparability.

Despite the potential benefits of providing access to microdata, NSOs face many challenges. They must consider the legal issues around microdata control and access, confidentiality/privacy concerns, and the costs associated with preparing public, restricted or scientific use files, as well as the cost to both the researcher and the NSO of maintaining secure research enclaves. Providing transnational microdata access expands the list of legal and organisational considerations and may increase the public's concern regarding the confidentiality of data collected by the NSOs.

This discussion will focus on the technical considerations involved in providing metadata to support the discovery of microdata files and their content, assessing the need and eligibility for access to restricted data as well as operationalising access and confidentiality constraints.

1.1 Motivations and concerns

One of the primary motivations for increased access to microdata is the need to research more fully the social and economic changes within and between countries. Several approaches have been taken to increase access while protecting the confidentiality of respondents. These range from the production of Public Use Files which limit the geographic and/or topical detail but are more broadly available than files held by Research Data Centres (RDCs) where individual users must be authorised for access, must use microdata in the strictly controlled setting of the RDCs, and must have the outputs of their research thoroughly reviewed for confidentiality issues prior to leaving the RDC environment.

An RDC may be "physical," providing access at a specified location where security – including scrutiny – can be most controlled, and therefore the detail of the data accessible to the researcher may be greatest. There is also a trend toward virtual facilities which allow remote access for accredited researchers using technology to control what a researcher may see during access and receive as output. Automation may be facilitated in the case of remote access because the options available to the researcher are more limited.

Other options such as scientific or restricted use files for the researcher community and synthetic data are also in use. Each of these options presents limitations to the types of research that can be done. From an NSO perspective, it is often appropriate and necessary to be able to support more than one of these access options. In order to be cost-effective it is beneficial if information requirements common to multiple channels of access can be managed in a single, standard manner.

1.2 Legal and organisational layers

Legal and organisational concerns are many and varied, particularly with regard to transnational access to microdata. Given that the frameworks to address the legal and organisational issues exist, the technological concern is one of relaying constraints

and processes in a structured way, so that systems can be constructed to operationalise the vetting of researchers for permission to access specified data files, the maintenance of multiple channels for access, and executing confidentiality processes to ensure releasable results.

1.2.1 Access

The goal of the technology is to implement these rules in an automated way to facilitate processing of access requests, reduce duplication of effort where possible, and limit the level of human intervention in the process to the most critical situations. This does not mean an acceptance of a single process in every case, but the ability to capture the access accreditation process in a consistently structured way so that the process rules can be used to drive the automated accreditation system. How closely NSOs align their individual rules and processes depend ultimately on their legal and organisational environments. However, the development of a system that can interpret rules and processes expressed in a common language can be used to manage the different paths where rule and process agreement cannot be reached.

From the researcher perspective access also involves having sufficient metadata available for a wide range of data files to determine whether access to the restricted data file is required and if obtained, whether the data will support the desired research objective. This information needs to be explored and compared across data sets from a broad range of sources. To do this effectively requires the use of either a common repository or a system which can query multiple repositories using a single common metadata language.

From an NSO perspective, in order to provide researchers with the detailed content information needed, it is necessary to use either a common metadata language to describe such descriptive and structural information about the data, or a language which can be efficiently and effectively "translated" to that common language. Benefits result if the common metadata language used to manage access is consistent with the language used to provide descriptive and structural information regarding the data. (This topic is explored further in 3.1.)

1.2.2 Confidentiality

Once a researcher has access to selected data files for a designated project, the primary issue becomes one of securing confidentiality for the subjects of the data. It is common to apply some confidentiality procedures to data – for example, removing direct identifiers -- prior to publication. However, if a researcher is working with a subset of the file and/or linking it to another source, additional checks may need to be run against the resulting data file (physical or virtual) to ensure confidentiality. These algorithms may be run during the creation process or on the resulting file. The algorithms used for these checks, and the triggers for their application, can be processed by the system if captured in a standard metadata structure.

2 Technical and semantic approaches to dealing with legal and organisational considerations

From a structural perspective, the enforcement of access and confidentiality constraints is a process consisting of triggers, rules, and decision points. Automating this process depends on being able to express each event in a set of unambiguous paths with clear rules for the information collected, decisions made, and actions taken. Clearly the greater the harmonisation among the NSOs regarding the specific details of these processes the easier and less confusing it is for the researcher. However, legal agreement and consistency of specific details are not requirements. Even within a single NSO some microdata will have tighter restrictions than others and those differences need to be enforced. The fundamental requirement is a common, structured, machine-actionable means of expressing the restrictions – whatever the restrictions may be in a particular instance.

2.1 Use of a metadata standard to enforce clear communication

The assumption is that a clear process can be defined which eliminates or at least limits the need for individual human review of access requests and the vetting of output for confidentiality concerns. The path through the process can vary by researcher type, data file, owner, etc., but the steps in the process can be executed using clear decision rules. Those rules, including definition of the information collected, triggers for choice of processing path, and evaluation of content (decision-making), must be expressed in a consistent manner to enforce clear communication and accurate execution. In short, the use of a common metadata standard for relaying both the content of the various review processes and the microdata content to the system is required.

2.2 DDI as a means to capture metadata

The Data Documentation Initiative (DDI) is a metadata standard focused on the capture, processing, management, and preservation of metadata for microdata and its resulting data structures in the social, behavioural, economic and related sciences. DDI has two development branches. DDI-Codebook (DDI-C), originally published in 2000, focuses on a single data file, capturing structural and contextual metadata regarding the purpose, organisational context, collection, processing, and structure of the data file. DDI-C has been used in many social science archives since its publication, but more recently, as a result of the success of the International Household Survey Network (IHSN) Microdata Toolkit developed at the World Bank, DDI-C has been adopted by NSOs in over 80 countries. The second development branch, DDI-Lifecycle (DDI-L), was published in 2008 and has experienced rapid uptake among organisations dealing with complex or longitudinal data collection processes, including NSOs. As reflected by its name, DDI-L focuses on the full life cycle of data from concept development and management through the development

of data collection instruments, data capture and processing, dissemination, preservation, and analysis. The DDI Lifecycle model was one of the models used as a basis for the General Statistical Business Process Model (GSBPM) and can be seen reflected in the structure of the top row of this model (Gregory, 2011).

Additionally, the aggregate structures within DDI-C and DDI-L are closely aligned with the Statistical Data and Metadata eXchange (SDMX) model, which is now being used by NSOs to document and exchange macrodata. Special care was taken in clarifying and tightening this alignment during the production of DDI-L, making output of an SDMX file from DDI-L metadata a direct process. What DDI-L provides to the SDMX model is the source and processing information on the creation of cell contents and dimension structures (Gregory, 2011) -- for example, the recoding of a microdata age variable expressed in single years to a table dimension expressed in 10-year age cohorts, or the derivation process for a specific indicator.

This high level of intentional interoperability means that a system designed around DDI-L could manage imports and outputs for both the major metadata standards currently in use within NSOs. Because DDI-L is intended to support the full life cycle of data and metadata, it contains many of the metadata structures needed by a system to support transnational microdata access.

DDI-L also has the capacity for full description of organisations and individuals in order to identify relationships such as ownership, process responsibility, and access management. Access rules are captured in a way that supports machine processing if needed. Embargo and access limitations can be attached to a data series, data file, or individual data item within a file. Different access criteria can be defined for different classes of researchers. Descriptions of individual researchers can be created to allow for easy identification of those who have already successfully moved through a vetting process for researcher approval.

A goal of DDI-L is to support a metadata-driven statistical process (i.e., concept management, questionnaire development, data capture, and data processing). Because of this goal, DDI-L captures processing activities as both descriptions and related code that can be directly used by a system to either initiate or run a specified process such as a confidentiality review. This combination of human- and machine-actionable information provides a solid foundation to document process.

3 Support for capture of metadata along the statistical process

A major theme of the strategic vision of the High-Level Group for strategic developments in Business Architecture in Statistics (HLG-BAS) is industrialisation and standardisation of statistics production, including the path forward. The HLG-BAS Vision notes, *Like any established industry, the production of official statistical*

information should have its own industrial standards…a necessary foundation for development and exchange of the means of production among the statistics producers (UNECE, 2011).

It is known that many NSOs consider metadata standards to support controlled access to microdata from the wider perspective of "industry standards" (Vale (ed.), 2011). The underlying concept of industrialisation of statistical production and the role of industry standards in supporting this are summarised below. In recent years there has been a strong trend across producers of official statistics at national and international levels to generalise, rationalise, harmonise, standardise, and modernise the statistical business processes required to produce statistics rather than focus on each survey as a separate "cottage industry" activity. The GSBPM, cited earlier, has been an important point of common reference in this regard (Engdahl, 2011 and Vale, 2011).

A process-centred view leads to a focus on "value chains" – processes working together in an integrated, interdependent, globally optimised manner to support needs. In the case of the industry of producing official statistics, various classes of information are both the core product (e.g., statistics), and the core raw material (e.g., data). The metadata required to "drive" the statistical production process – together with the statistical information required to be produced for external users as an output from the statistical business processes – is a critical component.

Technical standards for ensuring consistency and interoperability when defining and exchanging statistical information are an important aspect of industrial standards related to statistical production. The need for exchange of information extends not only to exchange of statistical information between organisations but also to the flow of information between sub-processes within the statistical business processes internal to an agency.

The HLG-BAS Vision focuses on industrialisation and standardisation across the community of producers of official statistics. Some NSOs are suspending judgement in regard to how quickly and consistently this is likely to proceed at an international level in practice. Nevertheless, many of those NSOs are seeking to apply strategies related to industrialisation and standardisation to their production activities at the national level.

While no one existing standard is commonly recognised as ideally suited for every purpose, technical standards for statistical information already exist (e.g., DDI-L and SDMX) and have been widely and successfully implemented to support a range of requirements. There is no evident support (or rationale) for creating a new, completely independent standard from first principles. There is, however, vigorous discussion about how SDMX and/or DDI-L might best be harnessed in an integrated and consistent manner across the industry. The ongoing SDMX/DDI Dialogue process is one example (UNECE/Metis, 2011).

3.1 Applying the concept of industrialisation and use of industry standards to enabling and managing microdata access

From an industrialisation perspective the microdata accessible via the channels discussed in 1.1 are one facet of the outputs which should be produced from a statistical business process. The metadata required to drive these sub-processes should be considered in the context of the information required to drive the broader statistical business process – targeting the maximum level of integration, efficiency, and reuse in terms of the information required

It is important that the common metadata language used to describe how rules and restrictions apply to microdata products and the variables within them is consistent with the common metadata language used to describe those products and variables for other purposes, such as driving statistical sub-processes that create and populate the product as well as the publication of documents to support access and use of the microdata products.

In considering standards as they exist today, Gregory et.al (2011) note:

- SDMX has some ability to describe microdata for exchange purposes, although this was not the primary use case for which SDMX was originally designed.
- DDI-L is specifically designed to support description of microdata in both human-readable and machine-actionable forms, where the latter supports automated processes related to the microdata.

Examples of differences between DDI-L and SDMX that are important to the microdata access scenario include:

- DDI-L supports structural description of variables.
- DDI-L supports structural descriptions of relationships between different types of microdata records (e.g., records for households and records for persons within each household). Such record relationships can be important when describing and applying confidentiality constraints for access to microdata.

While SDMX has particular structural strengths related to description and exchange of aggregate data, DDI-L currently provides more extensive structural capabilities related specifically to microdata.

The nature of the agreed future industry standard mechanism for harnessing these capabilities (e.g., direct use of DDI-L or representing semantically equivalent structural information using SDMX-ML “Metadata Structure Definitions”) is not critical to the current discussion.

It is useful to note that a number of NSOs (e.g., INSEE, Statistics New Zealand, and Australian Bureau of Statistics) have evaluated DDI-L, recognised its strengths, and

are in the process of exploiting its capabilities. Each of the three agencies listed intends to combine use of DDI-L to support statistical production with use of SDMX. The OECD Microdata Access Group has discussed development of a common industrialised approach to facilitating microdata access through a common system. The current intent is that the proposed access system would be based on DDI-L, which can also import content from SDMX and DDI-C (Thomas, 2011).

A number of data archives around the world have long used DDI in supporting management of and access to microdata. Currently, no all-encompassing industrialised microdata access capability exists within the data archive community as a whole. However, a number of useful frameworks, methods, and tools do exist which might be leveraged to develop an industrialised approach. A consistent, federated approach across NSOs and national data archives is likely to have benefits for both communities as well as for end users, who may find that some microdata relevant to their research is held by NSOs while other content is held by national data archives. This is, for example, a key consideration for the Data without Boundaries (DwB) initiative in Europe, which brings together research institutes, national data archives, and NSOs. Metadata standards to support "boundary-less" microdata access (particularly DDI and SDMX) are the focus of a specific work package within DwB.

3.2 Support through DDI

As described in 2.2.1 and 2.2.2, DDI-L already provides a range of capabilities relevant to supporting controlled access to microdata. These capabilities almost certainly do not yet encompass all the metadata (structured according to an industry standard common language) required to drive some of the fully automated processes related to rules and confidentiality that are aspirations for the future.

This is recognised by the DDI standards bodies and community, with a working group already initiated to identify additional requirements in this regard and to extend the support provided by future releases of DDI-L.

This is indicative of the nature of DDI as a responsive standard driven by practitioner needs. The generic release schedule for DDI sees two incremental releases of the standard each year, with an emphasis on extensions which preserve backwards compatibility. Unlike some software products, in practical terms there is not an issue with implementations of previous versions of the DDI standard becoming unsupported when new versions are released.

DDI-L therefore appears to offer an excellent starting point and framework for moving forward in terms of further developing an industry standard metadata language for supporting controlled access to microdata, integrated with the processing of microdata more generally.

4 Combined DDI-SDMX model

4.1 How these models work together

DDI provides a rich description of microdata sets, including information about how microdata are aggregated to produce tables. One approach to describing the full process of research or statistical production is to use DDI to support the documentation of the early life cycle stages, and then to use SDMX as a way of expressing the tabulated end result. It is not coincidental that some features of SDMX and DDI are very similar - both standards were designed to accommodate this combined application.

The ability to express DDI metadata in SDMX and SDMX metadata in DDI is considerable -- sufficient in fact to enable a cross-walk between the standards. This feature is used when the standards are applied in a combined fashion.

4.2 Relationship to GSBPM

One very typical case for this combined use of DDI and SDMX is in support of the GSBPM. At the national level, and for some supra-national agencies, the inputs to statistical production are microdata, and the outputs are aggregates, which will often be reported or disseminated in an SDMX format. We can understand the application of DDI within statistical agencies in this context: it supports many of the microdata-related processes for the earlier stages of the GSBPM, and SDMX is used in the later stages.

4.3 Supporting a transnational access system

Within the context of projects such as Data without Boundaries, where researchers are being given access to official microdata, we can see that microdata documented with DDI could be easily exposed within controlled-access environments. Rather than requiring two different sets of metadata -- one to support statistical production, and the other to support secure access -- a harmonised profile of the DDI metadata to support both functions could be defined. This would have the benefit of saving resources, and providing both data producers and researchers with a consistent and useful set of metadata and documentation.

5 Conclusion

DDI is in widespread use within secure data centres around the world, and notably in those which provide remote access. Within Europe, we see DDI metadata being used within several institutes for this purpose. Examples include the IAB, the statistical arm of the German federal employment agency, within the RDC based there. At IZA, an economic research institute based in Bonn, a remote execution environment is

available, with the documentation for the data managed using DDI. There are examples in North America as well. In the US, many different organisations are using a DDI-based management infrastructure provided by NORC for remote access. Called a Virtual Data Center, this model is also being deployed by the United Kingdom's Data Archive (UKDA). Within the Canadian RDC Network, DDI is being used to help with data management, and a next-phase development will include the creation of DDI-based tools to assist with managing disclosure risk. DDI is also an important component in a new project for the US Census Bureau's RDCs that is just now taking shape. In summary, DDI is positioned to become an increasingly popular standard for managing secure microdata in all types of environments, both within the domain of official statistics and for researchers, supplemented by SDMX in some official statistical organisations.

References

Engdahl, Jakob, Irebäck, & Holmberg, Anders. (2011). *Tentative anatomy of a new generation of IT-architecture to support GSBPM-processes.* UNECE, EUROSTAT, OECD Statistics Directorate, WP.4, 12 April 2011.

Gregory, Arofan. (2011). *The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes.* Open Data Foundation.

Gregory, Arofan, et al. (2011). *Contrasting DDI and SDMX Capabilities for the Description of Microdata.* Open Data Foundation.

Gregory, Arofan & Heus, Pascal. (2007). *DDI and SDMX: Complementary, Not Competing Standards.* Open Data Foundation.

Thomas, Wendy (2011). *Paris Microdata Group and DDI.* DDI Directions, Volume V, Number 1.

United Nations Economic Council on Europe. (2011). *Strategic vision of the High-level group for strategic developments in business architecture in statistics.* UNECE, Conference of European Statisticians. ECE/CES/2011/1.

United Nations Economic Council on Europe, METIS Group. (2011). *SDMX DDI Dialogue – Overview Page.* UNECE/METIS.

Vale, Steven. (2010). *Exploring the relationship between DDI, SDMX and the GenericStatistical Business Process Model.* DDI Working Paper.

Vale, Steven. (2011). *The Generic Statistical Business Process Model and its Implementation in Practice.* UNECE Statistical Division, WP5e (ppt presentation).

Vale, Steven (ed). (2011). *OECD Microdata Access Group.* MSIS Wiki.