Standardization and quality metadata of statistical survey
Comparison between (SDMX- DDI) Standards

Waleed Mohamed*
CAPMAS, Egypt
Email: ***kant2012xp@hotmail.com***

Abstract

The National Statistical Office in Egypt(CAPMAS) is committed towards producing high quality and relevant statistical information following the principles of neutrality, objectivity, professional independence, rationality and confidentiality to help all users get what they need from statistics and use it in making decisions. From this prospective this paper will discuss the Standards of Quality for metadata: (SDMX)" Statistical Data and Metadata Exchange" – (DDI) "data documentation initiative", Objectives of the (SDMX), Type of Metadata in SDMX ("structural metadata"-"Reference metadata"), DDI and SDMX in the statistical process lifestyle to make Comparison between of them With a focus on the Impact of (GSBPM) Generic Statistical Business Process Model in harmonizing metadata and quality, and the experience of CAPMAS(National statistical office in Egypt) in Documentation the statistical survey depended on standards (DDI)-Dublin core Initiative (DCMI) and what the important lessons we get it in this field .

Keywords : Standardization; metadata; quality statistical processes;DDI;SDMX; Documentation; GSBPM.

Introduction

Recently, two technical standards for statistical and research data and metadata have been receiving much attention. Particularly for those working with both micro-data and time-series aggregates, there can be some confusion as to the relationship between these standards, and questions about which may be more appropriate for use in a particular application or institution. This paper describes the basic scope of each standard, and provides some information which may help in making a decision about which of them is most suitable.

1.1 What is Statistical metadata

Metadata is often defined as data about data. It is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource", especially in a distributed network environment like for example the internet or an organization. A good example of metadata is the cataloging system found in libraries, which records for example the author, title, subject, and location on the shelf of a resource.

Statistical metadata is structured information about statistics. This includes information used for producing, disseminating, understanding, finding and (re)using statistics.

1.2 Standards of Quality metadata:

Metadata elements grouped into sets designed for a specific purpose, e.g., for a specific domain or a particular type of information resource, are called metadata schemes. For every element the name and the semantics (the meaning of the element) are specified. Content rules (how content must be formulated), representation rules (e.g., capitalization rules), and allowed element values (e.g., from a controlled vocabulary) can be specified optionally. Some schemes also specify in which syntax the elements must be encoded, in contrast to syntax independent schemes. Many

current schemes use Standard Generalized Markup Language (SGML) or XML to specify their syntax. Metadata schemes that are developed and maintained by standard organizations (such as ISO) or organizations that have taken on such responsibility (such as the Dublin Core Metadata Initiative) are called metadata standards.

1.2.1- What is SDMX?

An initiative, started in 2001, aiming at fostering standards for Statistical Data and Metadata exchange (SDMX). The SDMX sponsoring institutions are the Bank for International Settlements, the European Central Bank, Eurostat (the statistical office of the European Union), the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistics Division and the World Bank.

1.2.2- SDMX Global Conference 2009

The major SDMX event in 2009 was the Global SDMX Conference in January 2009 in Paris, hosted by the OECD. The conference dealt with the following issues: the SDMX standards and guidelines, the SDMX implementation projects and an outlook on further plans. Separate sessions on hands-on capacity building (training) were also organized.

A summary report on this SDMX Global Conference is attached in Annex 1. The key message coming out of this conference was: the SDMX technical and statistical standards and guidelines have reached a level of maturity now that is good enough for statistical organisations to use and implement them.

1.3- Communication and capacity-building

Communication and capacity-building activities on SDMX are a fundamental part of the SDMX initiative. They have been organised in a decentralised manner by all the SDMX sponsoring organizations. Some of the most significant actions are:

- The SDMX website (http://www.sdmx.org) which provides a single point of entry for all information on SDMX, ranging from the documentation on standards and guidelines to the downloadable software, together with announcements, events and information on implementation activities and DSDs. The SDMX User Guide and other tutorials are available via the website;
- The SDMX Global Conferences, which have been held in Washington (January2007) and Paris (January 2009);
- Training courses which have been organised or supported by the sponsoring organisations.

2- Metadata in SDMX

In SDMX, "structural metadata" are those metadata acting as identifiers and descriptors of the data, such as names of variables or dimensions of statistical cubes. Structural metadata must be associated with the data, otherwise it becomes impossible to identify, retrieve and browse the data. "Reference metadata" are metadata that describe the contents and the quality of the statistical data (concepts used, metadata, describing methods used for the generation of the data, and metadata, describing the different quality dimensions of the resulting statistics, e.g. timeliness, accuracy). While these metadata exist and may be exchanged independent of the data and its structural metadata, they are often linked (“referenced”) to data.

The idea is that it should be possible, using the SDMX standards, to exchange or share the data and the metadata that will allow a thorough understanding and interpretation of the corresponding statistical data.

2.1-objectives of the Statistical Data and Metadata Exchange (SDMX)

- create and maintain technical and statistical standards and guidelines to be used and implemented by the sponsoring or other organisations dealing with statistical data and metadata.
- improve efficiency by preventing duplication of effort. The SDMX standards and guidelines build on existing technical and statistical standards
- The SDMX standards also aim to ensure that appropriate metadata always come along with the data, making the information immediately understandable and useful. For this reason, standards for metadata exchange are extremely important in SDMX. At present, this part of the SDMX standards is only partially developed, but the plans for future editions of the standards comprise full development of metadata standards..

3- The Data Documentation Initiative (DDI)

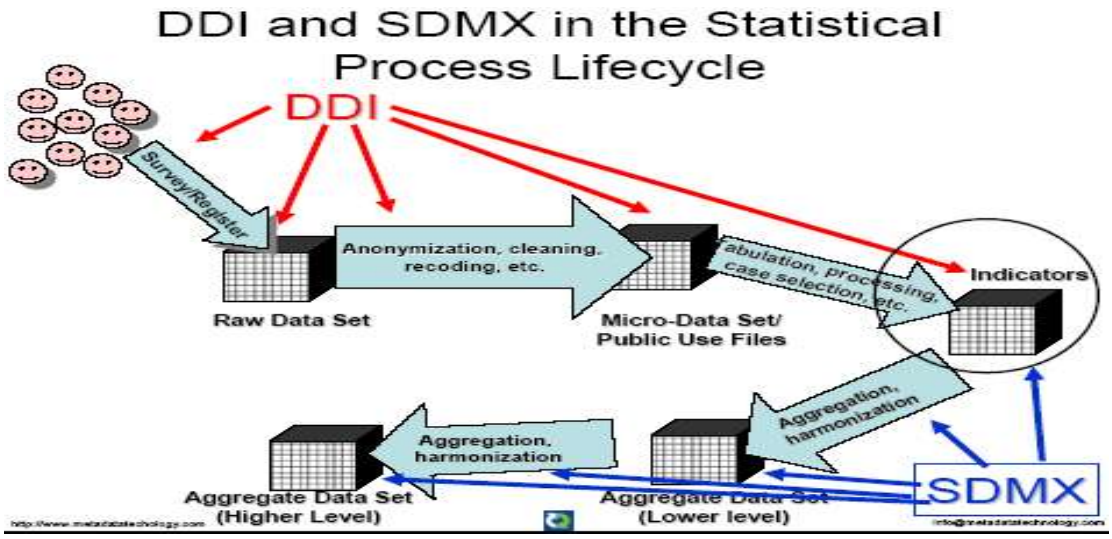
The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data. A membership-based Alliance is developing the DDI specification, which is written in XML.

By creating a consistent framework for micro data documentation, the DDI has the following features:

- Interoperability
  DDI-compliant documentation can be exchanged and transported seamlessly, and applications can be generically written, because the documents are homogeneous.
- Richer content
  The DDI provides data analysts with broader knowledge about data content, because the DDI initiative provides a comprehensive set of elements that can describe micro-datasets as completely and as thoroughly as possible.
- Multi purpose documentation
  A DDI codebook can be restructured to suit different applications, because it contains all the information necessary to produce different types of output.
- On-line analytical capability
  DDI documents can be easily imported into on-line analysis systems, rendering datasets more readily usable by a wider audience. This is made possible because the DDI markup extends down to the variable level and provides a standard uniform structure and content for variables.
- Search capability
  Field-specific searches across documents and studies are made possible, because each of the elements in a DDI-compliant codebook is tagged in a specific way.

4- DDI and SDMX in the statistical process lifestyle

SDMX and DDI take very different views of the data exchange process and data lifecycle.



One of the major cases for SDMX is the reduction of the reporting burden, both from the national level to the regional and international level, and among regional and international organizations. Often, the same data is reported many times by organizations to other organizations, and in each case a slightly different format for the data is required. SDMX addresses this issue by standardizing the formats, and by providing that the needed metadata accompanies the data. These bilateral data exchanges are often conducted using comma separated values (CSV) to format the data. Because CSV can be formatted in a large number of ways, it is not always possible to understand a data transmission without a specific knowledge of the format. Even if the formatting is evident in a similar

5- Comparison between SDMX&DDI

SDMX	DDI
SDMX provides XML formats for describing data and independent metadata structures, which can be user-configured to hold any concepts desired. They also provide XML formats based on these configurations. The concept of exchanging a data set or a metadata set is the primary focus in SDMX, which is optimized for the exchange of aggregate data. The typical case is the exchange of time series data.	DDI provides the ability to describe a rich set of metadata in an XML format, with an emphasis on micro-data, but also allowing for tabular formats and multidimensional cubes. In the 3.0 version, DDI supports all phases of the lifecycle from a description of concepts and the survey instrument used to collect data to the end product held in a data archive and used for analysis. DDI 3.0 also provides an XML format for micro-data and tabular/multi-dimensional data, but very often the data is held in text or statistical software specific binary files. The user-configurable aspects of DDI ("variables") are mixed with specific metadata fields.

6- Impact of Statistical Business Process Model in harmonizing metadata and quality " Generic Statistical Business Process Model" (GSBPM)

The term Business Process Model (BPM) is the noun form of Business Process Modeling and refers to a structural representation which defines a specified flow of activities in a particular business.

Business Process Modeling is performed in order to improve process efficiency, usually resulting in an improved quality as well.

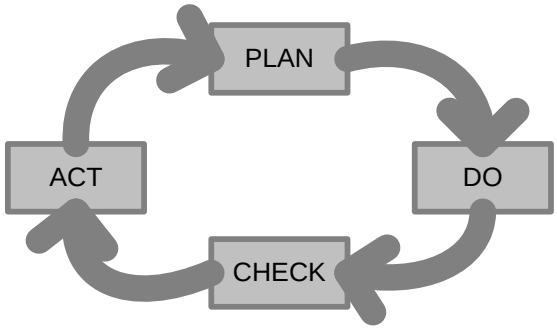
Quality management and quality assurance concepts have existed for a long time in the private sector, but for the public administration they have become an important issue since the time when the taxpayers increased their expectations for quality products and services in return for their money.

Concepts such as documentation, process description, metadata, quality assurance, and quality assessment were initially used in engineering and manufacturing, later followed by information and communication technology industry and software development in particular. Presently, statisticians have a very difficult task to adopt (or customize) the frameworks and models developed originally in the engineering and software industry, in order to make them relevant with regard to the particularities of the statistical production system where the main process is to deliver information to customers that they can use for their decisions.

The original intention when designing GSBPM was to provide a basis for statistical organizations to agree on standard terminology to aid their discussions on developing statistical metadata systems and processes, but later GSBPM was extended with the inclusion of the overarching processes (UNECE Secretariat 2009).

Plan-Do-Check-Act (PDCA) Cycle and how to standardize feedback loops

.



Question now is how to translate the PDCA cycle into the statistical business process model?

GSBPM starts with Specify needs and stops with Evaluate, but ideally it should be viewed as a system with feed back from Dissemination into Specify needs. Such a system could provide sound basis for identifying and linking process metadata and quality metadata.

The current GSBPM viewed from another perspective shows similarities to the waterfall model: progress flows from the top to the bottom, like a waterfall (Wikipedia 2010), except the occasional situations when some elements are forming iterative loops.

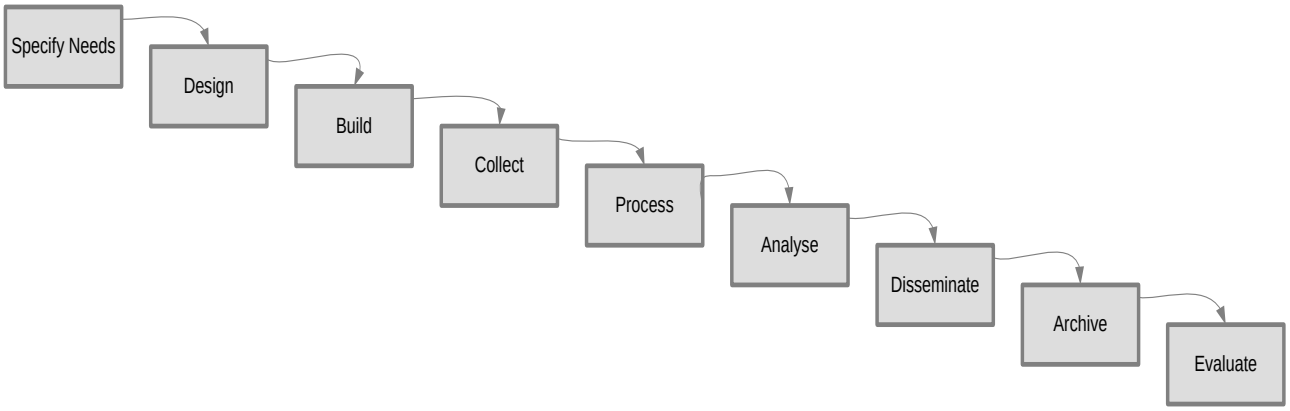


Figure 5: GSBPM viewed as Waterfall Model

The acceptable alternative of this model, the V Model (Wikipedia 2010), usually used in the software industry, could be applicable in managing the complexity of the statistical BPM. SSO's activities are focused on investigating the option to embrace it into the own business process model.

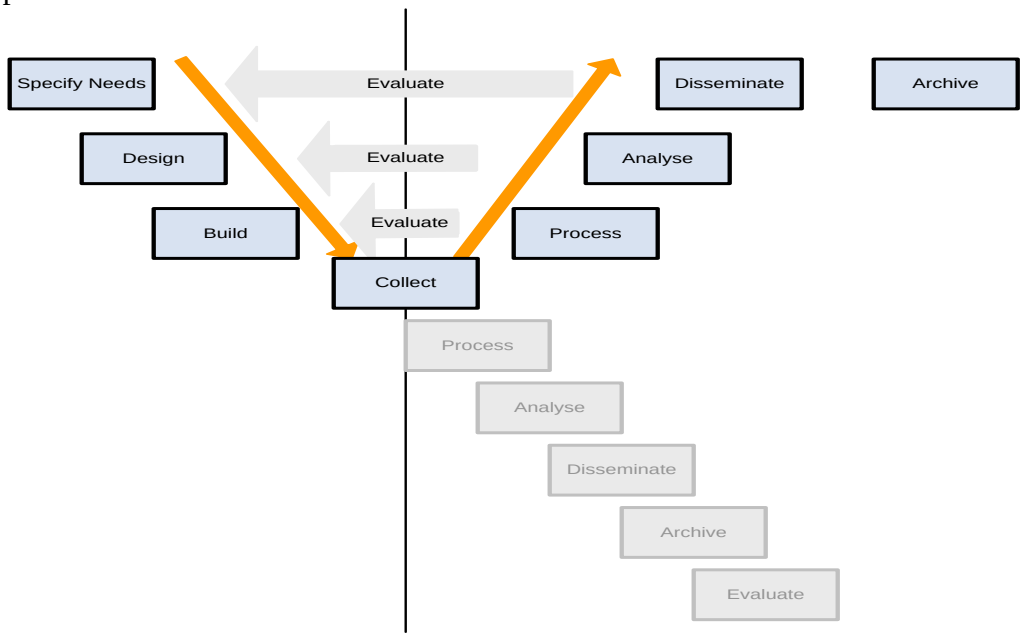


Figure 6: From the Waterfall Model to the V Model

In the above V Model, two major views of the metadata life cycle can be clearly identified:

- Left side (inputs): Activity-oriented view of metadata life cycle – description of processes
- Right side (outputs): Entity-oriented view of a metadata life cycle – description of deliverables.
- If the outputs do not meet the desired specifications, the process itself or the system behind has to be changed. A simple example is a survey with unacceptably high non-response rate. System changes may comprise changes in questionnaires or in the time

schedule for fieldwork. A more fundamental change would be to base the survey on other data sources for example.

- The V Model is lifecycle project process model providing a framework for process quality assessment and improvement. It is a variation of the waterfall model that makes explicit the dependency between planed activities and resulting activities. All activities from requirements to deliverables focus on permanent evaluation, whereas evaluation of the system is in the peer processes, in addition of the evaluating iterations of separate processes and sub-processes.

7-Documentation statistical survey based on DDI&DCMI"dublin core metadata initiative" using program(micro data management Toolkit) and the experience of CAPMAS(National statistical office in Egypt) in this field .

Documentation statistical survey based on DDI (micro data management Toolkit):

"From the archivist's and the end user's perspective, a "good" dataset is one that is easy to use. Its documentation is clear and easy to understand, the data contain no surprises, and users are able to access the dataset with relatively little start-up time."

The data documentation, or metadata, helps the researcher to:

- Find the data they are interested in. Without names, abstracts, keywords and other important metadata element it might be difficult for a researcher to locate the datasets and variables that meet his or her research requirements. Any cataloguing and resource location system - be it manual or digital - is based on metadata.
- Understand what the data are measuring and how the data have been created. Without proper descriptions of the design of the survey and the methods used when collecting and processing the data, the risk is high that the user will misunderstand and even misuse them.
- Assess the quality of the data. Information about the data collection standards, as well as any deviations from the planned standards, is important knowledge for any researcher who wants to know whether the data are useful for his or her research project.

Traditionally, data producers and data archives produced text-based codebooks. Today's alternative to text-based codebooks are XML-based codebooks, produced according to international metadata standards such as the Data Documentation Initiative (DDI) and the Dublin Core. To facilitate the documentation of microdata, the IHSN distributes the Microdata Management Toolkit, and promotes the adoption of international good practices.

Microdata Management Toolkit developed by the World Bank Data Group for the International Household Survey Network (IHSN) is designed to address the technical issues facing data producers. The aim in developing the Toolkit is to promote the adoption of standards for international microdata documentation, dissemination and preservation, as well as to foster best practices by data producers in developing countries. It complements other efforts by the

IHSN to produce and distribute tools and guidelines for improved management and use of microdata.1

The classic case for using DDI - especially for versions 1.*/2.*, but no less for version 3.0 - is the documentation of studies resulting from the administration of surveys. Population and agricultural censuses and household, enterprises and other sample surveys, all lend themselves to the use of DDI as an after-the-fact way that archives can document the metadata needed by researchers to make best use of the data. Such tools as the International Household Survey Network's (IHSN) Micro data Management Toolkit or the Nesstar software demonstrate how the metadata collected around a study can enormously improve navigation and understanding of the data collected.

The Toolkit comprises two modules. The Metadata Editor is used to document data in accordance with international standards. The CD-ROM Builder is used to generate user-friendly outputs (CD-ROM, website) for dissemination and archiving.

Steps of using toolkit in documentation.

1- With metadata Editor import data from various standard format (SPSS, STATA, ASCII, others), and provide comprehensive metadata in user-friendly screens.
   Data and metadata become one entity, saved in a single file.
   Notes: the free Nesstar Explorer program allows users to view metadata and re-export data to various common formats.
2- Generate various "metadata diagnostic" reports and automatically produce detailed survey documentation in PDF format.
3- With the CD-ROM Builder, produce a user-friendly html-output for sharing and preserving your data and metadata.

Total statistics that have been documented (published internal / external Dissemination) at CAPMAS 2010

	Name of the statistical	Dissemination of an internal	Dissemination of external
1	Annual Bulletin of the combined research workforce.	2008-2010	2007-2009
2	Annual Bulletin of Statistics of marriage and divorce.	2009-2010	2008-2008
3	Annual publication of statistics of births and deaths..	2007-2009	2008
4	Annual Bulletin of Statistics Industrial production in the private sector.	2007-2009	2008-2010
5	Annual Bulletin of Statistics industrial production facilities in the public sector	2008-2009	2010
6	Statistics form the basic electronic indicators to measure the information society, the family.	2007-2008-2010	2009
7	Statistics of education in the institutes are not subject to the ministries of	2006-2007	2007-2008

	Education, Al-Azhar		
8	Count activity hotel and tourist villages in the sectors of public and private	2008-2009	2010
9	Statistics of the building and construction companies to the public sector / business	2008-2009	2007-2008
10	Monthly Summary of Foreign Trade data - December	2008-2010	2009
	Total = 32 survey		

Conclusion

The Statistical Data and Metadata exchange (SDMX) initiative was launched in 2001 by seven organisations working on statistics at the international level: the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistical Division (UNSD) and the World Bank. These seven organisations act as the sponsors of SDMX.

The stated aim of SDMX was to develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an IT architecture for exchange of data and metadata. Organizations are free to make use of

whichever elements of SDMX are most appropriate in a given case.

With the Internet and the world-wide web, the electronic exchange and sharing of data has become easier and more common, but the exchange has often taken place in an ad hoc manner using all kinds of formats and non-standard concepts. This creates the need for common standards and guidelines to enable more efficient processes for exchange and sharing of statistical data and metadata. As statistical data exchange takes place continuously, the gains to be realised from adopting common approaches are considerable both for data providers and data users.

It should be clear from the discussion here that DDI and SDMX are standards which are related, but which are not in competition. They are very different in scope: where DDI is aimed at solving problems with the documentation of research, and across the micro-data lifecycle, SDMX is concerned with creating efficiencies around the exchange of aggregate data. DDI comes from the world of 17 data archives and social sciences researchers; SDMX springs from the world of official statistics.

REFERENCES

1- DDI and SDMX: Complementary, Not Competing, Standards , Version 1.0, July 2007

2- IHFAN Quick Reference Guide for For Health Facility Assessment Data Archivists, DRAFT - Version 2008.02

3- SDMX 2.1:" Comment Log to the Version 2.0 Specifications" 1 December 2010 available at : (http://www.sdmx.org).

4- Sdmx-userguide-version2009-1 available at : (http://www.sdmx.org).

5- Statistical Commission Forty-first session 23 - 26 February 2010:" Progress Report on SDMX" Prepared by the World Bank

6- www.surveynetwork.org/toolkit