STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

ECE/CES/BUR/2007/FEB/2/Add.1
29 January 2007

CONFERENCE OF EUROPEAN STATISTICIANS

For discussion

Third meeting of the 2006/2007 Bureau
Geneva, 12-13 February 2007

Item 1a of the Provisional Agenda

COORDINATION OF SUBJECT-MATTER DOMAIN GUIDELINES INVOLVING SDMX

Note prepared by the SDMX sponsors[1]

BACKGROUND

1. At the 8th meeting of the Committee for the Coordination of Statistical Activities (CCSA) in Montreal on 4-5 September 2006, SDMX Sponsoring Institutions[2] were asked to spell out envisaged steps towards coordinating activities of those national and international statistical agencies interested in developing subject-matter domain guidelines for exchanging data and metadata within the SDMX framework.

2. Better understanding of these steps was seen as providing a basis for CCSA's recommendation to the UN Statistical Commission in March 2007:

- to commend the goals and developments of SDMX in fostering standards and guidelines for data and metadata exchange;
- to recognize the importance of a wide use of SDMX standards for the exchange of statistical data and metadata among all participants in the international statistical system (national statistical offices, government agencies, central banks, international organizations, etc.);
- to encourage participation by national and international statistical agencies in their further development and use;
- to invite the SDMX Sponsoring Institutions to use the SDMX website to link to work of specific statistical domains that develop subject-matter data structure definitions and thereby provide some coherence to these efforts;
- to ask the SDMX Sponsoring Institutions to regularly report to the Commission on the initiative's progress.

3. The goal of this note is to outline the process by which those interested can participate and contribute to domain-specific developments, enabling greater efficiencies and reduced reporting burden in data and metadata exchange.[3]

---

[1] The report has been jointly prepared by the Bank for International Settlements (BIS), the European Central Bank (ECB), the Statistical Office of the European Communities (Eurostat), the International Monetary Fund (IMF), the Organization for Economic Cooperation and Development (OECD), the United Nations Statistics Division and the World Bank.

[2] Currently these are: BIS, ECB, Eurostat, IMF, OECD, UN, World Bank.

[3] Points articulated in this note have benefited from an exchange of views on the topic that took place during a meeting of the Bureau of the Conference of European Statisticians (CES) on 19 October 2006 in Washington, D.C. Participants included national statistical agencies along with international institutions, some of whom

EXISTING PROCESS TO ADOPT SDMX STANDARDS

4. The SDMX initiative provides a set of building blocks for the exchange of statistical information involving technical standards (an information model, exchange formats and an architecture) and content-oriented guidelines (cross-domain concepts, a list of subject-matter domains and a metadata common vocabulary).

5. The way in which SDMX standards and guidelines are developed and officially endorsed entails the following steps:

- the SDMX Secretariat, in consultation with experts working in Sponsor organizations and other experts in the field (including those working in other international organizations, national statistical offices and national central banks), develops draft standards and guidelines;
- the SDMX Sponsors Committee authorises the publication of draft technical standards and guidelines that appear on its website for public comment periods and for reviews of issues in meetings with their constituencies (especially national and statistical agencies and central banks);
- following this public consultation, the sponsoring institutions are then ready to follow an approval process for the release of its technical standards and those content-oriented guidelines that are across domains (i.e. cross-domain concepts and vocabularies). In addition, the technical standards go through an approval process of the International Organization for Standardization (ISO) for Technical Specification 17369 (SDMX).

6. This process has proved to be efficient and fully transparent, also considering that each organization, according to its own rules, has regularly informed its constituency about SDMX developments and that SDMX Sponsors as a whole reported to the UN Statistical Commission.

COORDINATION OF SUBJECT-MATTER DOMAIN DEVELOPMENTS

7. SDMX envisages three essential elements in developing subject-matter data structure definitions and code-lists in a transparent and coordinated way, building wherever possible on available cross-domain concepts and the metadata common vocabulary:

- a list of domains for which data structure definitions should be developed;
- organizations (or groups of organizations) developing data structure definitions for statistical domains (or groups of domains);
- a mechanism to foster coherence in data structure definitions (and associated code-lists) developed by different organizations/working groups in order to facilitate their use worldwide.

8. A preliminary list of statistical subject-matter domains already represents one of the SDMX content-oriented guidelines.[4] This list makes extensive use of the latest version of the United Nations Commission for Europe (UNECE) Classification of International Statistical

---

are sponsoring organizations of the SDMX initiative. A follow-up discussion is scheduled to take place at the CES Bureau meeting in Geneva in February 2007.

[4] The preliminary drafts of SDMX Content-Oriented Guidelines can be downloaded from the SDMX website http://www.sdmx.org.

Activities.[5] As a starting point, Annex 1 offers an initial high-level list of these domains; and Annex 2 provides a brief introduction to what's involved in a data structure definition.

9. The development of data structure definitions (and associated code-lists) for specific subject-matters requires two types of expertise: statistical and technical. While there is a longstanding co-operation on a worldwide level to develop statistical concepts and methodologies in a coordinated way (through various international working groups and task-forces), experience to develop data structure definitions and code-lists according to SDMX content-oriented guidelines is currently centred within the SDMX Sponsoring Institutions and a number of central banks and national statistical agencies. The Sponsors Committee sees the incremental development of domain-specific data structure definitions (and associated code-lists) as building on the work of both the domain specialists and SDMX-experienced experts. Broad-based collaboration among institutions and statistical experts has to be ensured, especially to:

- foster good practices for the development of sensible domain-specific terminology for concepts and code lists;
- facilitate awareness of important issues and possible mapping principles that may be necessary with respect to existing classification schemes and systems of countries and international institutions.

10. In this context, some supplementary information is provided in Annex 3 about the role of SDMX data structure definitions in creating efficiencies. Annex 4 also highlights kinds of implementations where the SDMX framework can be usefully applied.

COSTS AND BENEFITS

11. Recent experiences in developing data structure definitions have focused on external debt, monetary statistics, international financial statistics, balance of payments, commodity trade statistics, national accounts, short term indicators, agriculture data and some additional macro-economic statistical series. A well-established use of the framework has been in place for many years for a variety of statistics at the European System of Central Banks (ESCB), Eurostat and the BIS. Some data structure definitions, eg for ESCB and Eurostat, are also enshrined in EU legal acts.

12. Out of recent and earlier experiences, some understanding has been formed about the order of magnitude of estimated costs to develop the formal SDMX structures that are exchanged. If the classification schemes of the data sets or questionnaires are already known and well structured, including code lists, it may take only a few weeks of actual work to develop the basic components of the SDMX data structure. However, the time for needed for fine-tuning of the structure, by all parties reviewing the work, is estimated to take several months. The actual technical implementation of the data structure definition, using available SDMX tools or in-house techniques, is estimated at only a few days. The fewer the institutions/groups involved and the smaller the data sets, the shorter the actual and elapsed times. In addition, major components of data structure definitions that are developed can be re-used to address specific institutional requirements for national or international agencies.

[5] See the current version at http://unece.unog.ch/IntPres/disa.explorer.asp?Search=PAPE&Year=2006

13. In the case of external debt, for example, the data structure definition used for the BIS-IMF-OECD-World Bank Joint External Debt Hub (JEDH) took about three weeks of actual work over a time period of about six months. In the case of the OECD's National Accounts World Wide Exchange (NAWWE) project, the estimate for work associated with creating the data structure definition for the entire Eurostat-OECD questionnaire was about six weeks over a period of about twelve months. In both cases, actual technical implementation of the data structure definition took a few days.

14. In terms of benefits, there are clear advantages in building lists of concepts and associated code lists that can be used by a wide range of users for the exchange of statistical information. In principle, two institutions exchanging information can agree on the terminology between themselves and take advantage of SDMX technical standards. Similarly a disseminator of statistics can use its own specific terminology and SDMX standards. However, a more common set of content-oriented guidelines will facilitate data exchange and dissemination greatly as these can be shared by many institutions and users. It also avoids duplication of efforts in developing and maintaining lists of statistical concepts and associated code lists. The benefits can materialise most quickly in those areas where there already exist national and international statistical nomenclatures.

15. It should be noted that there is no need for individual institutions to change their internal content-oriented guidelines when common guidelines are developed and made available. Interfaces can be developed to map how data are identified internally with how they are identified during exchange or dissemination. Before SDMX, such interfaces had to be build for every single bilateral data exchange (or within closed groups of exchanging institutions). The more SDMX can promote common subject-matter domain guidelines, the fewer such interfaces will be needed in the future, or the simpler they will become.

16. In the case of the JEDH mentioned above, the launch of the technical application became straightforward once the data structure definitions for the external debt data were approved by the Inter-Agency Task Force on Finance Statistics. These definitions were easily deployed within the data contributed by the participating institutions as well as within the application created within the World Bank’s exisiting Development Data Platform and the connectivity to an SDMX registry infrastructure hosted at the OECD. The NAWWE project is also making use of SDMX and envisages implementation of common data structure definitions through a rapid-development approach that can provided incremental functionality over time.

17. As the methodological and computer-related work are part of the usual activities taking place at national and international institutions concerning statistical definitions and exchange processes, any specific trade-off between costs and benefits may be seen as associated with the specific nature of internal systems or interfaces to international systems. As the common formats of the SDMX structures provide "machine-readable" data, these yield opportunities at the technical level to lower costs (eg reduced manual processes such as re-keying of data or re-formatting). In addition, important efficiencies come from using a common approach by many institutions for many users, for example, exchanges with several international organizations and availability via websites.

OPEN PROCESSES

18. A central factor for the success of the proposed approach to creating domain-specific data structure definitions is the transparency of the process to develop them, in particular to ensure the involvement of users. Therefore draft domain-specific concepts, data structure definitions and code lists should be issued for public comment and review before their final approval by the respective working groups/experts that will develop them. SDMX Sponsors are in the process of studying how the SDMX website could be used to assist this process, especially in the context of providing links to the work done on subject-matter domain sites and thereby offering an opportunity to foster coherence in the process.

19. The following steps are envisaged:

- when an organization (or a group of organizations) decides to develop data structure definitions for a specific subject-matter, it would communicate this intention to the SDMX secretariat, using a pre-defined template made available on the SDMX web site. The information would be made public on the same web site and disseminated to CCSA members and UNSC members;
- the organization (or group of organizations) would develop the data structure definitions for their specific subject-matter domain ensuring an open consultation period for public comments; the SDMX secretariat can make suggestions that the organizations(s) cooperate with other organization(s) that may be working on similar activities concerning the same or related statistical domains. Subject to availability of resources, the SDMX Secretariat might also assist in reviewing notes (or completed templates) provided by subject-matter domain groups that indicate how their work is conformant with the SDMX framework, in order to ensure coherence of their work within the broader scope of SDMX developments;
- when the final data structure definition is ready, it would be posted on the SDMX website for public use. This way the SDMX web site would become an evolving repository where standard data definitions and code-lists developed for specific subject-matters would be made available to all interested parties. The UN Statistical Commission would receive regular updates in order to be informed of progress in these developments.

20. The process would not require any change in existing coordination and governance arrangements in countries and international institutions. SDMX will only play a coordinating role. The success of the proposed approach will depend on the goodwill of the various institutions and groups involved to share their expertise and resources and to cooperate effectively.

ANNEX 1. Preliminary SDMX list of statistical subject-matter domains

Domain 1: Demographic and social statistics	Domain 2: Economic statistics	Domain 3: Environment and multi-domain statistics
1.1 Population and migration	2.1 Macroeconomic statistics	3.1 Environment
1.2 Labour	2.2 Economic accounts	3.2 Regional and small area statistics
1.3 Education	2.3 Business statistics	3.3 Multi-domain statistics and indicators
1.4 Health	2.4 Sectoral statistics	3.3.1 Living conditions, poverty and cross-cutting social issues
1.5 Income and consumption	2.4.1 Agriculture, forestry, fisheries	3.3.2 Gender and special population groups
1.6 Social protection	2.4.2 Energy	3.3.3 Information society
1.7 Human settlements and housing	2.4.3 Mining, manufacturing, construction	3.3.4 Globalization
1.8 Justice and crime	2.4.4 Transport	3.3.5 Indicators related to the Millennium Development Goals
1.9 Culture	2.4.5 Tourism	3.3.6 Sustainable development
1.10 Political and other community activities	2.4.6 Banking, insurance, financial statistics	3.4 Yearbooks and similar compendia
1.11 Time-use	2.5 Government finance, fiscal and public sector statistics	
	2.6 International trade and balance of payments	
	2.7 Prices	
	2.8 Labour cost	
	2.9 Science and technology	

ANNEX 2. What is a data structure definition?

1. Statistical data is represented with numbers, such as “17369”. If you are presented with a number - as above - you can't tell from the number alone what it is a measurement of. A number of questions come immediately to mind: What is the subject of the measurement? What units does it measure in? What country or geographical region, if any, does it apply to? When was the measurement made? The list of questions is potentially endless.

2. Behind each of these questions there is a particular "concept", which is used to describe the data. In our questions above, these descriptor concepts are Subject, Unit of measure, Country, and Time. The simplest explanation of a Data Structure Definition is that it is “a set of descriptor concepts, associated with a set of data, which allow us to understand what that data means”. There is more to it, however.

3. There is a higher level of package known as a "Data Set". Typically, it is maintained and published by an agency, so that it becomes a known source of statistical data. A basic structure is emerging: we have Observations, grouped over time into Series, which are grouped into Data Sets.

4. In order to be able to exchange and understand data, a Data Structure Definition tells us about the relevant dimensions and what the possible values for each dimension are. This list of possible values is known as a code list. Each value on that list is given a language-independent abbreviation - a "code" - and a language-specific description. This helps us avoid problems of translation in describing our data: the code can be translated into descriptions in any language without having to change the code associated with the data itself. Wherever possible, the values for code lists are taken from international standards, such as those provided by ISO for countries and currencies.

ANNEX 3. Greater efficiency through SDMX data structure definitions

1. When SDMX standards and guidelines are used to put together a data exchange among two or more partners this involves the specification of a “data structure definition” (also known as a “key family”). Some of the concepts and related code lists used in the actual exchange are cross-domain (e.g. unit of measure, frequency), while others are domain-specific (e.g. financial instruments, aggregates of national accounts).

2. A data structure definition specifies a set of concepts which describe and identify a set of data. It tells us which concepts are the dimensions (identification and description) and which are attributes (just description). It also tells us which code lists provide the possible values for the dimensions and attributes or if these values are of another data type such as numeric or free text. For example, the data structure definition for the BIS-IMF-OECD-World Bank Joint External Debt Hub (JEDH) contains the concepts that identify and describe the external debt data required for the JEDH website. It contains concepts that uniquely identify the data, such as the creditor country, the debtor country, the financial instrument, the maturity of the instrument, the data frequency, etc. For each of these concepts, a code list is provided and must be used when reporting the data. The data structure definition also contains concepts that describe the data reported, such as currency unit and decimal, which only describe the data, as the same data could be reported using a different number of decimals.

3. SDMX is focused on exchange or sharing between institutions and dissemination to users. Two institutions can agree on the exchange terminology between themselves and take advantage of the efficiencies that follow from the use of computer-readable formats that are provided through the SDMX framework. Interfaces can be developed at institutions to map how data are identified internally with how they are identified during exchange. But building lists of concepts and associated code lists that can be used by a wide range of users for the exchange of statistical information allows greater efficiencies in transmissions and web access to a wide range of data and metadata.[6] Commonalities make it easier to identify what is being exchanged and to navigate catalogues of information that make use of these identifiers.

4. For more general use by many institutions (and associated greater efficiencies) a more common set of content-oriented guidelines can facilitate the exchange. This is achieved by using common cross-domain and domain-specific terminology, like in the case of JEDH. Dissemination of data and metadata on the web also follows the same principle: greater gains from efficiency follow from greater commonality in use of content-oriented guidelines for what is contained in the exchange. Supplementary metadata can provide explanations of how "local" practices for describing and compiling the data may be different from the "common" terminology used in exchanges. Institution-specific terminology can also be used in exchanges, hopefully building on common high-level domain-specific content-oriented guidelines.

5. For example, to develop the JEDH website the participating organizations have already defined “code lists” for several dimensions, such as country, time, unit, etc. These dimensions can be made available on the SDMX website to be eventually re-used by other organizations

[6] A new section of the SDMX website will contain information on available concepts and code lists and how they are used in constructing data and metadata structure definitions.

for bilateral or multilateral exchanges of data pertaining to other subject-matters (price indices, national accounts, etc.). This approach would minimize the costs for those institutions who want to use SDMX guidelines. Therefore, the target is a framework for interfaces between as many "owners" of statistical data and metadata as might be interested over time in making use of this exchange framework.

*** * * * ***

ANNEX 4. Going forward with implementation of SDMX standards

1. Efforts in collaboration in subject-matter domains that follow the SDMX framework are taking shape in national accounts, balance of payments, external debt, financial statistics, education, agriculture, population and some indicators associated with the Millennium Development Goals. In particular, actions that foster the use of the SDMX framework, and therefore require the development of data structure definitions and code-lists, include the following:

- *SDMX for batch exchange*: international organizations already use SDMX formats for batch data exchange with national institutions for a wide range of statistics and these processes are being extended, facilitating more efficient data flows that can be handled by internal systems at sender and receiver ends of the exchange.
- *SDMX joint hubs*: international organizations that already have joint data collections are thinking about the development of joint hubs, possibly under the auspices of existing international working groups. For example, Eurostat, OECD and UNESCO already use a joint questionnaire on education statistics and share data collection and verification work, and data are used to produce a joint publication. They already discussed to move to SDMX standards and guidelines both for data collection and data dissemination (joint hub), developing data structure definitions for all variables covered by the exercise.
- *SDMX for existing data collections*: organizations who already collect data through electronic questionnaires using various text formats are transforming their questionnaires in SDMX format. This can be done individually or in a more co-ordinated way, both through bilateral contacts with other agencies active in the same field, or involving international working groups active in that particular statistical domain. Of course, when joint data collection activities are carried out by two or more IOs an agreement on data structure definitions has to be found among participants, wherever possible adopting and building on those already existing.
- *SDMX for dissemination*: SDMX formats are already used to disseminate data, irrespective of how they have been collected.

* * * * *