STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

Joint UNECE/Eurostat Work Session
on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 11
English only

Topic (i): Infrastructure issues for statistical metadata

Common Open standards for the Exchange and Sharing of Socio-economic Data and Metadata: the SDMX Initiative

Prepared by the Bank for International Settlements, the European Central Bank, the International Monetary Fund, the Organization for Economic Co-operation and Development, the Statistical Office of the European Communities and the United Nations Statistical Division

Contributed paper

Common Open standards for the Exchange and Sharing of Socio-economic Data and Metadata: the SDMX Initiative

I. Introduction

> "The BIS, ECB, EUROSTAT, IMF, OECD, and UN have joined together to focus on business practices in the field of statistical information that would allow more efficient processes for exchange and sharing of data and metadata within the current scope of our collective activities. The goal is to explore common e-standards and ongoing standardization activities that could allow us to gain efficiency and avoid duplication of effort in our own work and possibly for the work of others in the field of statistical information."

1. This quotation is from the statement that was distributed prior to a Workshop on Statistical Data and Metadata Exchange that was sponsored by the above institutions and held at the International Monetary Fund in Washington, D.C. on September 6–7, 2001. More than 100 participants from all regions of the world attended the meeting. At the concluding session of the workshop the participants recommended that the sponsoring institutions lead an international endeavor resulting in the creation of the standards envisaged in the quoted statement.

2. Immediately following the September workshop, the sponsoring institutions met to initiate the process of creating a framework to respond to the recommendations made at the workshop. They agreed to formalize a task force to address Statistical Data and Metadata exchange (SDMX).

3. Part II. of this paper provides an historical perspective for SDMX. Part III. identifies the growing need to exchange data and metadata. Part IV. sketches the requirements for exchange standards, while Part V. identifies the relevant technologies. Part VI. suggests how a standards creation process may be organized.

II. A brief history of standards

4. International attention to the topic of standardized cross-national statistics dates back at least to the League of Nations, which held the International Conference Relating to Economic Statistics in 1928. In the post WW II period, standardization was carried forward with the issuance of *Measurement of National Income and the Construction of Social Accounts* by the United Nations in 1947 and the *Balance of Payments Manual* by the International Monetary Fund in 1948. These documents provided standard definitions of statistical concepts, and work on these and a variety of other statistical topics has continued to the present.

5. The advent of commercial computing in 1953 led to the development of internal standards for coding statistical data. However, it was the advent of inexpensive electronic communications in the last quarter of the twentieth century that led to the development of standards for electronic exchange of information. This occurred first in the commercial world with the Sabre airlines reservations system and the SWIFT network for banking transactions.[1] The public sector stepped into the arena with the publication of the "Guidelines for Trade Data Interchange" (GTDI) by the UN/ECE in 1981, which led to ISO 9735 Electronic Data Interchange for Administration, Commerce and Transport (EDIFACT) syntax rules published in 1988.

6. In the early 1990s, the syntax for an EDIFACT message called Generic Statistical Message (GESMES) was developed. This led to the implementation of BOPSTA (a GESMES type message) in the mid-1990's by EUROSTAT, the IMF, and a limited number of their member countries. A new GESMES profile called GESMES/CB was introduced in 1998-99 by the Bank for International Settlements, the European Central Bank and EUROSTAT (and adopted by the IMF). By the turn of the millennium, electronic exchange of statistical data had become a standard business practice

[1] The truly pioneering *Sabre* system went on-line in 1960 and the Society for Worldwide Interbank Financial Telecommunication (SWIFT) initiated transactions in 1977. See www.sabre.com/about/index2.html?b=1&a=history and www.swift.com/index.cfm?item_id=1243

among these central agencies and their member countries.

7. While the above was taking place, an alternative to EDIFACT, which involved a different form of exchange, was also in the making. This part of the story begins with the issue of ISO 8879: *Information processing – text and office systems – Standard Generalized Markup Language (SGML)* in 1986. SGML was developed to address the difficulties of moving text into formatted (photocomposed) documents in a generalized and reusable manner. A derivative of SGML, called Hypertext Markup Language (HTML), was developed together with the World Wide Web (WWW) by scientists at CERN.[2]

8. HTML, a non-proprietary derivative of SGML, is used to control the layout of web pages on computer screens. HTML's strengths lie in its ability to format text, graphics, and links to other text etc. in an environment of overlapping pages on a computer screen and in its ease of use. It became, and remains, one of the driving technologies of the Internet.

9. As the amount of information on the web exploded, the need for a markup language that addressed the content of information embedded in text began to be recognized. To meet this need, the World Wide Web Consortium (W3C) created the Extensible Markup Language (XML) initiative in May of 1996. The result of this initiative was the publication of version one of XML in February of 1998.[3]

10. The power of XML is that it structures the information contained in text or associated with data and metadata.[4] This structure allows information to be found within the body of text without doing a full text search. It also allows the exchange of information in an unambiguous manner. The power of XML was quickly recognized by the information processing industry. Today XML products and standards abound.

[2] HTML was used to create the original web site at CERN. The general release of the WWW on CERN computers occurred in May of 1991. See www.w3.org/History.html and public.web.cern.ch/Public/ACHIEVEMENTS/web.html

[3] See www.w3.org/Press/1998/XML10-REC

[4] See www.w3.org/XML/1999/XML-in-10-points for a summary of the basic concepts of XML.

III. The need for data and metadata

11. New needs for economic data on a cross-national basis coincided with the above history. The economics of general equilibrium and emergent Keynesian macroeconomics, which implied that whole economies could be managed, generated a need for macroeconomic data. In addition, the lessons of the great depression of the 1930's lead to the understanding that economies need to cooperate if a more stable world economy was to be achieved. These events also drove the development of statistical methodologies.[5] The need for increasing volumes of macroeconomic data that were definitionally comparable across economies became the conventional wisdom of national and international economic managers and market participants.

12. These events also defined the need for a new type of standardized information. This information consisted of comprehensive descriptions of who, what, where, when, and how national data are produced and disseminated..

13. An example of this new form of information about the data is the OECD *Quarterly National Accounts: A report on the sources and methods used by OECD Member Countries* (1979). The IMF began developing comprehensive frameworks for macroeconomic metadata for the Special Data Dissemination Standard (SDDS), which was established in 1996. This was followed by the introduction of the General Data Dissemination System (GDDS) in 1997. [6] EUROSTAT introduced Euro indicators, a collection of data and metadata covering the euro-zone and EU-15 in 1999, in the wake of the new European Monetary Union.[7] In early 2001, the Euro indicators were pulled together into a single web site, where metadata are shown in the SDDS format. Many countries have also developed their own web sites containing a mix of data and SDDS or GDDS metadata.

[5] A list of statistical methodologies is located at http://esa.un.org/unsd/progwork (see *Methodological Publications in Statistics*)

[6] See dsbb.imf.org.

[7] See europa.eu.int/comm/euroindicators.

14. During the 1990's, the work undertaken within the UN/ECE work sessions on statistical metadata (METIS) produced a significant consensus on some conceptual issues and more specific guidelines, such as the "Guidelines for Statistical Metadata on the Internet". Statistical metadata were defined as "data which are needed for proper production and use of the data they inform about"; data describing statistical data and – to some extent – processes and tools involved in the production and usage of statistical data.[8]

15. Following the pattern for data, the newly developed sets of metadata are also being exchanged between and among national states, regional and international organizations, and the general public. The need for standardization of metadata exchanges is a logical outcome of the increasing need to exchange metadata.

IV. The scope of the SDMX initiative

16. The scope of SDMX initiative is, in general terms, the exchange of data and metadata within the collective activities of the SDMX organizations. Therefore, the activity is currently limited to the topical ground of socio-economic statistics. This section covers many of the core business issues relating to the exchange of this statistical information.

Business models for exchange

17. Two distinct paradigms for the exchange of statistical data and metadata have emerged. The first paradigm it that of direct exchange of files between parties who have made prior arrangements for the exchange. The second paradigm involves the placement of data/metadata on a web site that then can be selected by consumers using efficient tools and processes.

18. The first of these models may be described as the partner – hub model, named to describe the typical relation between the parties. In this model the partners all ship sets of data/metadata to a central collection authority (the hub). At a national state level, the partners are the economic units in an economy and the hub is a national authority responsible for the particular type of data/metadata being collected. At the international level the partners are member states and the hubs are international or supranational organizations such as the BIS, ECB, Eurostat, IMF, OECD and UN. In this model the principal responsibilities for the information producer are to prepare the data/metadata and to initiate the transaction. The data/metadata receiver is the more passive participant, waiting for the information to be sent.

19. The second exchange model has been described as the dissemination model. In this model a data/metadata producer places the information on a site that is accessible to data/metadata consumers. The consumers then access the site and read the information. In this model the transaction is initiated by the information consumers that pick and choose what data/metadata they want. With the advent of Internet technology, the site of choice has become a web site[9].

20. Many international organizations and national agencies already have on-line databases available to external users. Because the design and content of these databases vary enormously, there is wide variation in the ability of such on-line facilities to meet user requirements. Furthermore, the evolution of such databases and their creation by other agencies will mean that data will become even more accessible. This trend highlights the need for organizations to make metadata even more available. Unfortunately, experience to date is that the provision of metadata with data significantly lags the availability of data.

21. Both of these models will continue to be actively used. Each has clear advantages in specific contexts. The first is more suited where the data requirements of users are "stable" over long periods of time, the second where requirements are either ad hoc or subject to frequent change. The business requirements of both models need to be addressed.

22. In both models there is a need to design metadata content standards in parallel with the data exchange

[8] See UN Statistical Commission and UN/ECE publications "Guidelines for the Modeling of Statistical Data and Metadata", United Nations, Geneva, 1995 and "Guidelines for Statistical Metadata on the Internet", CES Statistical Standards and Studies, n° 52, Geneva, 2000.

[9] A special case of the dissemination model is where data consumers poll a number of data producers for a specific piece of information that is needed.

standards. Designing standards in this way would allow metadata to be used more effectively than is now possible to compare national methodological practices.[10]

Data and metadata models

23. One question that arises when speaking of standards for data and metadata exchange is whether data and metadata should be taken together in one standard. Alternatively, should different exchange standards be developed for data and metadata. In order to address this issue, we need begin to look at how data and metadata are used (i.e., the business models for data and metadata).

24. Pure data is barren. For example, the game scores 4 to 3 and 2 to 1 mean almost nothing until you identify the sport, the team names, and when the games were played. The data are 4, 3, 2, and 1. The metadata (information about the data) provided is that these data are game scores. The metadata needed for the data to be useful are the sport, team names, and dates. It would also help if it were explained that these were women's Olympic soccer (football) games.

The point of the example is that all data comes with a substantial amount of metadata, and that these data and metadata are inseparable. That is, neither is very useful without the other.

25. However, there is another type of metadata which can stand alone when separated from the data and make good sense.

26. Examples of this metadata are the information in the OECD's sources and methods publications[11] and the information about national data systems of a country found on the IMF's Dissemination Standards Bulletin Board (DSBB). The information in these publications defines how data on a given topic may best be organized into a structure of component parts and how it is to be or was compiled. None of these publications contains any data.

27. Given that we have at least two different ways of approaching data and metadata, it appears that we may need two different standards for their exchange. One standard would describe data and its associated metadata. The second standard would describe metadata that resides in some form of catalog.

Time series and tabular data

28. There is yet another fundamental way to differentiate classes of data that are commonly used in socio-economic statistics. These classes are time series data and tabular data.

29. Working with data where each observation is associated with a particular span or point in time has its own set of problems. A time series is a collection of observations on the same phenomenon where all the time signatures are either points in time or spans of time. With time series you must deal with which type of time, points or spans, definitions of the calendar you are using, and social conventions applied to that calendar (e.g., what is the work week). Macroeconomic data are typically expressed in time series.

30. For tabular data one needs to define the dimensions of the matrix and the logic of the breakdowns along each dimension. Some of these dimensions may not be numeric (e.g., the race of the head of household or the existence of running water, electricity, indoor toilets, etc. in the household). Census data is typically presented in tabular form.

31. SDMX would begin with an attempt to develop common standards. However, the different approaches may follow different business rules and there may therefore be a need for separate models for data and metadata that are time series and data and metadata that are tabular.

V. The technologies of the standards

32. The title of this section uses the plural in both of its nouns. Earlier, the paper outlined the need for a number of standards. It is also the case that different standards are likely to use different technologies. Moreover, as technological innovation may be expected to continue to move forward, new standards will need to be developed in order to attain the advantages offered by the newer technologies. At present, there is a need to address at least two technologies that are applicable to statistical data and metadata exchange. These are the

[10] See *Developing a Common Understanding of Standard Metadata Components: A Statistical Glossary* at http://www.unece.org/stats/documents/2002.03.metis.htm.

[11] See www.oecd.org//mei (refer *National Methodological Practices*)

GESMES specifications of UN/EDIFACT and the Extensible Markup Language (XML) specification standard of the World Wide Web Consortium.

33. The technologies used need to comply with three technical principles. These are:

- the structure should be captured in a standard way so that it can be used by any tool or technology and not be dependent on a specific vendor's product;
- the structure should be described in a language that is extensible, allowing for additions as new information is created; and
- the language used to describe the structure should be independent of formatting and presentation features, thus allowing these features to be determined by each user.

EDIFACT

34. The EDIFACT technology facilitates the construction and interpretation of messages containing statistical data and associated metadata. EDIFACT is very compact and highly suitable for fully automated, repetitive data exchanges. These messages can be self-contained and logically complete. A perceived weakness of the EDIFACT message format is that it takes considerable effort to set up EDIFACT based exchanges, so that it is not well suited for ad hoc exchanges. It would also be unsuitable for exchanges that arise out of browsing a collection of web sites and picking up pieces of data here and there.

XML

35. XML is far less compact (though compression techniques may take care of this) but well supplied with commercially developed tools and more appropriate for data sharing over the web. XML is extensible, platform independent, and supports internationalization and localization[12]. XML-based messages are self-contained and logically complete; they can be human readable and they are also well suited for small ad hoc data exchanges.

VI. The standards creation process

The open process

36. The sponsors of the SDMX initiative endeavor to focus on the creation of common standards that will suit the needs, not only of themselves, but also of their member states and their data user communities. A general view is that there is a need to create an open and transparent process for participation of member states and data/metadata consumers in the development of the standards. However, the specifics of the implementation of this view are complex. They are still under discussion by the SDMX sponsors. As expressed in the literature on this topic, the idea of an open process centers on a few key principles.[13] They are as follows:

- all parties interested in engaging in the effort to create a given standard and willing to provide their own time and effort may participate;
- the cost of participation should be born by the participants;
- the cost of participation should be minimized to the extent that it is not a significant barrier to willing participants;
- the intellectual property developed by the process should be freely available for public use at no cost;
- the process should be governed by a formal democratic process; and
- the deliberations taking place within the process should be archived and publicly visible.

37. The SDMX initiative intends to use these ideas as guidelines for the process it intends to employ in facilitating the development of standards for data and metadata exchange. By doing so, it is expected that barriers to the sharing of the intellectual property developed by SDMX will be minimized. In addition, these ideas are intended to encourage the widest possible adoption and to encourage the marketplace to

[12] See *XML in 10 points* at www.w3.org/XML/1999/XML-in-10-points

[13] See *A Scalable Process for Information Standards* at www.xml.com/pub/a/2001/01/17/oasisprocess.html

develop products that support usage of the standards created.

SDMX Work Program

38. This paper has suggested some important topics that could be within the scope of the SDMX initiative, in synergy with other already existing groups. These include the following:

- Time series data with metadata;
- Tabular data with metadata;
- Metadata catalogues, glossaries, dictionaries, etc.;
- Partner – hub exchange models;
- Dissemination exchange models;
- EDIFACT syntax based implementations; and
- XML syntax based implementations.

39. It was agreed at an early stage among the SDMX sponsors that the initiative would build, as much as possible, on existing data models and message structures. This, of course, is not an easy task. Different organizations have good reason to protect the investments they have already made. Existing working groups, task forces, and committees have their respective mandates and procedures to be respected and accommodated. The global setting adds complexities. Thus, a concrete work program, with assigned tasks, is still being discussed among the sponsors, as is a formal organizational structure.

40. In order to support its work program, SDMX has created is own web site www.sdmx.org and e-mail address SDMX@imf.org. This web site now includes all the presentations from the September SDMX Workshop and information about contacting SDMX partners. It is expected that the work on a number topics will be initiated in 2002. These activities will be announced on the web site together with any relevant mailing lists to keep their participants and observers informed.

41. SDMX solicits all statistical agencies and all persons involved in reporting to or using the data produced by these agencies, who have an interest in participating in any part of the work of SDMX, to contact SDMX at the above e-mail address and express their interests, business requirements and priorities.