

**Economic and Social Council**Distr.: General
3 June 2011

English only

Economic Commission for Europe

Conference of European Statisticians

Fifty-ninth plenary session

Geneva, 14-16 June 2011

Item 3 of the provisional agenda

Organization of data collection and sharing, and the management challenges for the implementation of Statistical Data and Metadata eXchange**The use of Statistical Data and Metadata eXchange in Istat****Note by Istat, Italy***Summary*

While there is a tremendous pressure on the resource of the statistical organizations which face everyday new data demands without a parallel financial allocation, on the other hand the development of information and communication technology has been encouraging the production and publishing of “competitive” statistics from many different institutions and enterprise. In this scenario synergies, standardization and optimization of the process and infrastructure are the only solution to this challenge, and Statistical Data and Metadata eXchange can be considered an essential “building block”.

Istat has decided to invest in the modernization of its statistical information system, consequently the Statistical Data and Metadata eXchange Single Exit Point will be one of the dimension in which this modernization will be developed.

I. Introduction

1. The aim of this paper is to illustrate the challenges that Istat is facing in order to modernize its information system and the progress in Statistical Data and Metadata eXchange (SDMX) implementation activities as part of its strategic vision. In particular the work conduct has been to exploit the standard for improving the business process behind the life cycle of statistical data and for improving the harmonization of the contents.

II. The current context and the future scenarios

2. Istat has always followed with great interest the development of technologies that have a potential impact on the collection, processing and dissemination of statistical information: this interest has enabled an ongoing evolution in the instruments that it uses, with significant gains in efficiency and effectiveness. In recent years the Institute has made appreciable progress in satisfying new demands for statistical information, while seeking at the same time to reduce the burden on respondents and optimise productive processes.

3. However new challenges have been raising for Istat in the last few years, due to the development of Information and Communication Technology (ICT) that has implied not only to opening new technological scenarios for the statistics but also to reducing drastically the cost of producing data, enabling ever larger numbers of institutions to publish statistics from administrative sources and from sample surveys.

4. The information systems of large and medium-sized enterprises churn out a continuous stream of data, sometimes released more frequently and with fewer constraints than the information provided by national statistical institutes (NSIs).

5. Very often those data are immediately analyzed and in turn transformed into complex indicators. Google has recently started producing price indices based on online transactions. Others use data from credit card purchases to produce indicators of sales of goods and services in real time.

6. Practically, via Internet, we have access to a boundless mass of quantitative information produced all over the world. Visualization techniques offer extraordinary communication tools which allow even nonexperts to understand the meaning of complex phenomena, summed up in a few numbers.

7. On the other hand there is a tremendous pressure on the resource of the statistical organizations which face everyday new data demands without a parallel financial allocation: the obligation to reduce costs while improving efficiency and quality has become one of the main issues that all the statistical organizations have to face.

8. At the same time, confusion among analysts, and the public opinion in particular, grows with the number of available sources and public statements which cite contradictory data about the same phenomena.

9. If, then, this is the world in which we live and it is reasonable to expect that some of the trends noted here will continue, and perhaps accelerate, we should wonder how “official statistics” should evolve in order to perform that essential role that it is called upon to play for the functioning of our society.

10. In this context the modernization of the statistical information systems is one of the main issue that each organization has to face and, furthermore, the NSIs, in their role as the main coordinators of the national statistical systems are called on to initiate and, often, coordinate the transformation. The adoption of standards, the harmonization of the

information (concepts, classifications and dictionaries), the overcoming of the “stovepipe” logic which still today characterizes many statistical processes and new methods for interchange of data and metadata are all sustainable answers to this changes.

III. Stat2015: Istat strategy for the next years

11. In the last twelve months Istat has launched a multi-year project called “Stat2015”, which is set to plan and bring about a deep technological and organizational leap for an Institute which aims to be a leader in the processing of information for statistical purposes and in relations with respondents and users of statistical information (individuals, households, enterprises and institutions).

12. In short, the project can be broken down into three key dimensions:

A. Collection and initial processing of data and metadata

13. In this case, it will be developed a consistent system of information acquisition methods oriented towards digital data capture, controlling, monitoring and verifying the quality of processes, timeliness, reducing the statistical load, integrating data from other sources, as well as the sharing of infrastructures and the reusability of information within the National Statistical System. A goal to pursue therefore is the integration of the eXtensible business reporting language (XBRL) protocol¹⁰ (which has become the standard for the documentation and interchange of financial flows, and is already used for the transmission of financial statements and the “Unico” tax return forms) and the SDMX protocol (for statistical data exchange), both based on the eXtensible markup language (XML) language. The construction of a statistical taxonomy in the XBRL is underway and could be applied starting from the most burdensome surveys for enterprises, such as those used to produce the data required by the European Community Regulation on structural business statistics.

B. Management of the information

14. The aim is to create a system based on methods of processing, managing and archiving data and metadata oriented towards quality control of the information produced, transparency and reproducibility of results, as well as conserving and protecting the information heritage. In order to achieve these aims it is necessary to shift away from the “stovepipe” logic which still today characterizes many statistical processes. In this respect Istat intends to further develop, for statistical purposes, the integration of microdata derived from administrative and statistical sources.

C. Dissemination and communication

15. For several years Istat has published its statistics through traditional channels in common with many statistical institutes, such as printed and online publications, websites, etc. Specifically, online dissemination has been either through files containing predefined tables in various formats or through databases. For example, through Istat’s website it is possible to access several databases which not only differ from each other in terms of ways of access and navigation, but also provide statistics in an inconsistent and/or partial manner (with data not always accompanied by adequate metadata).

16. In order to improve the service for users and prepare the Institute to interact with them in a radically different manner, a number of actions have been instituted. They include in particular:

(a) The new website: having undergone a complete redesign both graphically and in terms of navigation, and available in Italian and English, it will make information much easier to use, even for the least expert users. By plugging in applications for data visualization and querying, and through real-time dialogue with open systems of dissemination, the new site will enable direct access to numerical information, interconnected with accompanying meta-information. The information will be arranged in pages made up of modular information elements and offer two main types of access: a public tab for non-authenticated users (with contents determined by the need to provide everyone with a fixed set of information and to pursue specific communicative strategies) and a customized tab for authenticated users, that is, an aggregator of contents with tools which the user can select from a list, configuring the site in accordance with his/her own needs (MyIstat);

(b) The new I.Stat data warehouse: available in Italian and in English, I.Stat is set to become the single container in which all statistical macrodata regarding the various economic, social and environmental phenomena will be stored. I.Stat (the Italian version of the equivalent OECD.Stat) presents both individual datasets in a simple form, according to a predefined presentation scheme, and also allows the user to define the tables corresponding to his/her needs, to save them and then update them with a single click, so as to always have the most up-to-date data to hand;

(c) New press releases: revised in form and content, press releases on individual topics become “portals” themselves: through them, with just a click, the user has access to complementary information content (including databases) connected to them;

(d) New graphical visualisation tools: based on Statistics eXplorer, they enable (starting from data stored in I.Stat) the static and dynamic display of phenomena of interest through graphs and maps, which can then be included on other sites, so that they are enhanced with comments and other information (photographs, documents, etc.) produced and selected by the user;

(e) The new SDMX Single Exit Point: this important step towards a new form of dissemination and communication of statistical information is just the first result of the ground-breaking transition to a way to unlock data warehouses and open data and metadata to those users that need to interact in machine-to-machine manner.

IV. Statistical Data and Metadata eXchange in Istat

17. Istat has been working on SDMX since 2004. At the beginning, the participation to pilot projects (SDMX Open Data Interchange, Demography Rapid Questionnaire, Census Hub, Euro Group Register) launched by Eurostat within the European Statistical System (ESS) has allowed to get knowledge and experience in order to evaluate its usage beyond the data exchange. In fact beside the policy for data exchanging, the use of SDMX gives an opportunity to integrate heterogeneous information systems and to harmonize the statistical contents.

18. Also the *vision*¹ “on the production method of EU statistics: a vision for the next decade”, shared at ESS level, goes in this direction. In particular Eurostat has stressed² that

¹ <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF>

one of the areas where the *vision* may have particular implications for dissemination will involve the existing “Single Entry Point” at Eurostat as well as a “Single Exit Point” in each NSI, fed by a data warehouse. Logically, such a data warehouse would handle not only data intended for Eurostat, but also data which are disseminated to the general public by the NSI. A common Single Exit Point for dissemination and as well as for reporting to Eurostat or other international organizations would save work and would help to ensure the coherence of national and European statistics.

19. As part of the Stat2015 vision and in the context of a new data dissemination policy, Istat has decided to put in place a series of activities related to the SDMX implementation within the Institute itself and, more in general, with the National Statistical System:

(a) The installation of I.Stat data warehouse (based on OECD.Stat data warehouse), to foster a single repository for all dissemination datasets proceeding from the statistical production directorates. In this context all the already existing dissemination databases will converge into I.Stat;

(b) The development of a server of classifications to give full and easy access, to all internal and external users/applications, to harmonized classifications also in SDMX format;

(c) The installation of a SDMX infrastructure, based on the SDMX-RI developed by Eurostat and extended by modules developed by Istat. This infrastructure will act as a Single Exit Point for data and metadata (SDMX Registry) allowing a machine-to-machine dissemination/reporting for SDMX structural metadata and data through a web service.

20. In order to streamline the work plan, several supporting actions have been taking place:

(a) Development of the SDMX Istat Framework³;

(b) Participation to the ESSnet⁴ on SDMX phase I and II;

(c) Internal training for statisticians and IT staff (three introductory sessions in 2010, one session on data modelling in 2011);

(d) Active participation to the development of the new 2.1 version of SDMX;

(e) Set up of a task-force composed by statisticians and IT staff with the main aim of modelling datasets through the definition of all the necessary SDMX structural metadata. Furthermore this task-force is in charge of the mapping of concepts and codes stored within the I.Stat and other dissemination databases, with concepts and codes defined in the DSDs.

² See Dissemination Working Group, agenda point 5 – Luxembourg 2009.

³ The Istat SDMX Framework will be integrated, during the ESSnet on SDMX phase II, within the SDMX-RI developed by Eurostat.

⁴ European Statistical System Centres and Networks of Excellence is an instrument, created by Eurostat, in order to find synergies (from cooperation between partners), harmonization and dissemination of best practices in the ESS. Participating countries receive financial support under the form of a Multi-beneficiary Grant Agreement.

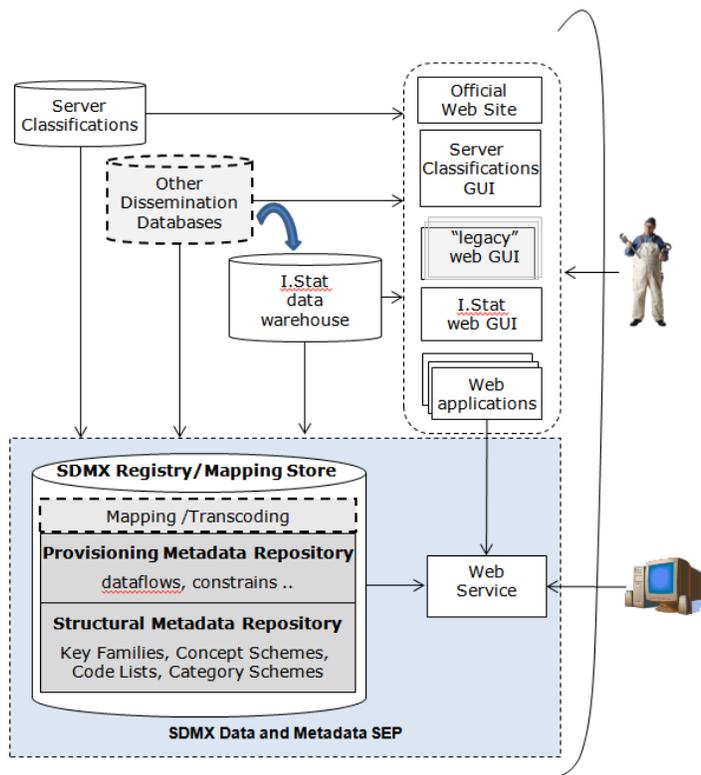
V. The Statistical Data and Metadata eXchange Istat Single Exit Point

21. The SDMX Single Exit Point (SEP), besides allowing an easy access to data in formats that makes it easy to reuse (machine-to-machine modality), will:

(a) Reduce the existing dissemination “stove pipes” to publish data one, and let the counterparty “pull” data and metadata as required. The aim is to minimize the number of repository and the number of formats (excel, csv, etc.) in which the data is currently disseminated;

(b) Improve quality and efficiencies by promoting a) the use of consistent statistical concepts and code lists across domains, b) preservation of the meaning by coupling data with metadata that defines and explains it properly, c) exchanging and presenting data and metadata in an open, enduring format (“XML”) rather than a proprietary one.

22. The figure below gives a schematic representation of the technical implementation of the SDMX infrastructure put in place.



23. At the base of this SDMX infrastructure there is the idea of “mapping” between “legacy” and SDMX metadata. In this way it is not necessary to change the structure and the content of the metadata within the already existing repositories. In order to facilitate the mapping processes, that otherwise it would be time-consuming and very expensive, a suitable tool, named Mapping Assistant, was created.

24. The Mapping Assistant is a desktop tool that facilitate the mapping between the structure metadata provided by an SDMX-ML Data Structure Definition (DSD) and those that reside in a dissemination database of an NSI dissemination environment. Through a graphical interface a user can create the mapping between dissemination databases and DSDs. The Mapping Assistant was designed to edit and store the mapping information in a Database Management System (DBMS) called Mapping Store, and communicates with both the Mapping Store and the dissemination databases in standard Structured Query Language (SQL).

25. The SDMX Registry/Mapping Store is the central building block of the infrastructure. It acts as a “back-office” application for internal and external access for retrieving SDMX structural metadata. It stores mappings between the SDMX and the native concepts and code lists and, finally, it takes track of which dissemination database contains the data that a user/application can request through the Web Service.

26. For the time being the Registry is able to store the following SDMX artefacts: category scheme, concept schemes, code lists, key families, data flows. Constrains can be queried on-the-fly.

27. The Web Service is the building block responsible for exposing the data and metadata using a Web Service interface that provides SDMX-ML messages. In this context it implements two interface with the following methods:

(a) For querying datasets: GetGenericData, GetCompactData, GetCrossSectionalData;

(b) For querying structural metadata: QueryStructure.

VI. Conclusions and lessons learned

28. After 10 years SDMX is mature enough to move from a tactical to strategic perspective. Therefore it is time to move away from pilot exercises and to consider SDMX as an essential component of any modern statistical business architecture: SDMX not only for guaranteeing the progress in “data sharing”, but for facilitating and improving the integration, harmonization and interoperability.

29. Consequently the involvement of the top and middle management is a basic pre-requisite.

30. SDMX is not only for IT staff, but statisticians are the subjects that have to drive the progress in the right direction, and with the right speed.

31. Capacity building actions are essential for starting any SDMX project.

32. Last but not least, re-using software and experiences not only reduces costs for implementing SDMX architecture or running SDMX projects, but allows to achieve objectives in a short period.