IAOS Satellite Meeting on Statistics for
the Information Society
August 30 and 31, 2001, Tokyo, Japan

Influence of the Internet on data collection and dissemination in the European Statistical System (ESS)

KNÜPPEL, Wolfgang
KUNZLER, Uwe
Eurostat, Luxembourg

Abstract

This paper has three main parts. Part 1 examines electronic data reporting technologies and some concrete projects in this area within the ESS, including electronic questionnaires and intelligent Web forms. Part 2 analyses the current situation of EDI standardisation: the transition from EDIFACT to XML and the adoption of ebXML for the ESS. Part 3 deals with electronic dissemination of statistics. A recent benchmarking on this topic has shown some interesting results. Further issues are usability testing and the use of agent technology for a global data access network.

A glossary of acronyms and abbreviations used in this paper and a short list of references are included; Web addresses within the text point at further information sources.

Keywords

European statistical system, Electronic data reporting, Standardisation, Electronic dissemination, Internet, Web forms, EDI, EDIFACT, XML, ebXML, Agent technology

Introduction

The European Statistical System (ESS) can be seen in different ways:

- **what is it for**: production of statistics for and about the European Economic Area.

- **who are the partners**: close co-operation of national authorities and a central EU authority.

- **how does it work**: system of agreed standards, tools and organisational methods.

National authorities are the National Statistical Institutes (NSIs) and other bodies (e.g. central banks, customs authorities, …) responsible in each Member State for producing European statistics. The EU authority is Eurostat, the Statistical Office of the European Communities in Luxembourg.

The statistical process incorporates three major phases: collection, production and dissemination. This paper examines data collection and dissemination. Both are based on the exchange of information i.e. on communication: raw data are communicated between data provider and data collector; dissemination is communication between producer and consumer of statistics.

The Internet – more precisely the World Wide Web, or ‘the Web’ for short – has changed the way communication takes place. Of course this affects also the world of statistics. Statistical data collection and dissemination becomes more and more electronic and automated. There are several reasons for this phenomenon, including:

- **user demand**: users want data on the Web which they can download and process; enterprises want electronic questionnaires making declaration faster and thus less expensive.

- **political pressure**: there are many political initiatives at national and trans-national level aiming to bring government on-line; examples in Europe are ‘e-Government’ or ‘e-Europe’.

- **technical progress**: who would have expected the Internet revolution 10 years ago? New technologies arise every day – people and businesses use them, the statistical world has to follow.

In the ESS, raw data are mostly collected at national level, following the EU principle of subsidiarity. The national authorities produce and publish national statistics, but they also forward collected or aggregated data to Eurostat. Eurostat’s task is the creation and dissemination of European level statistics.

This system requires a close co-operation between the partners and agreed standards, especially in the field of Electronic Data Interchange (EDI). This is why a separate chapter of this paper (part 2) treats EDI standardisation.

Part 1: Data collection

Data collection in the ESS comprises two phases: raw data collection – which is mainly done in the member states by national authorities – and transmission of (usually aggregated) data to Eurostat. This chapter concentrates on raw data collection.

The collection of raw data includes direct (or primary) and indirect (or secondary) data collection. Primary data collection is based on surveys with data providers (or their agents). Some statistical offices claim that they collect more than 90% of their raw data through secondary collection, for example from taxation or customs databases. However, this paper will not examine secondary data collection.

EDR – Electronic Data Reporting

Direct raw data collection based on electronic questionnaires is also known as Electronic Data Reporting (EDR). EDR stakeholders are: data providers (individuals, households, enterprises, administrations, …); data collectors (national statistical institutes, national banks, Eurostat, …); software suppliers (offering commercial tools for statistical data reporting); and standardisation bodies (providing the required standards).

Standardisation will be examined in part 2 below. A problem is the involvement of commercial software suppliers. Ideally, they should produce electronic questionnaires or integrate statistical modules into their existing products like business management or accounting systems. There were a number of pilot and research projects on that at national and European level in past years. The success, however, was quite limited. There are not many acceptable commercial software solutions for statistical reporting. The reason may be that software houses do not see enough potential to earn money in this field, or it might be that they think it is too difficult with the current standards (i.e. EDIFACT).

There are two basic EDR technologies: Computerised Self-Administered Questionnaires (CSAQ) and Web forms. CSAQ systems existed already before the Internet revolution. A CSAQ is a stand-alone software package that is installed on the data provider’s workstation. It allows statistical data to be captured and packed in a format accepted by the data collector. Today’s CSAQ systems make more and more use of the Internet: data are sent over the Internet; program updates can be downloaded from the data collector's Web site; long code lists may be consulted on the collector's web server; and so on.

A successful example of a CSAQ is IDEP/CN8, an electronic questionnaire for Intrastat declaration; Intrastat is the statistical system relating to the trading of goods between EU member states. IDEP/CN8 was developed and is maintained by Eurostat. It is available in most EU member states in the respective languages. The competent national administrations take care of distribution and user support. Currently, IDEP/CN8 is used by over 50,000 enterprises. The existence and success of IDEP/CN8 is the result of a close and fruitful co-operation between the relevant partners within the ESS.

While traditional CSAQ systems like IDEP/CN8 cover only one statistical survey, multi-questionnaire CSAQ systems can be used for several surveys in parallel. New questionnaires can be added whenever needed, for example through download via the Internet. EDISENT, one of the first multi-questionnaire systems, was developed in the TELER research project (1996-1999). A more recent development is now being used in Austria (e-Quest, opened to the public in April 2001).

Web forms

Web forms are electronic questionnaires that do not require software installation on the provider's computer. A Web form is available on the data collector's Internet server and can be accessed and completed with a usual Internet browser. This kind of solution is offered more and more by statistical offices in most EU member states. As an example you may take the UK Intrastat Web form (available since 1998; see http://www.hmce.gov.uk – Intrastat) or the German w3stat system (available since 1999; see http://www.w3stat.de).

The IQML research project (2000-2003) goes a step further: the Web form becomes intelligent. An intelligent questionnaire is able to interrogate enterprise databases and to extract the required data. This will become possible through the use of XML and metadata standards. The IQML project will develop a software suite and support the development of an XML standard for intelligent questionnaires based on the Common Warehouse Metamodel (CWM) of the Object Management Group (OMG). The IQML software suite will include tools for questionnaire design, questionnaire presentation, database interrogation, survey administration and most importantly a metadata repository. More information under http://www.epros.ed.ac.uk/iqml/.

Part 2: Standardisation of EDI

In general terms Electronic Data Interchange (EDI) means the exchange of information by two computer applications in a structured way without human interaction. This is mainly needed to automate business-to-business (B2B) communication, e.g. to send supply orders, bills and so on in an automated way.

EDI can also be applied in the statistical world. Two statistical institutions may exchange aggregated data, or a provider of raw data (e.g. an enterprise) may send data to the collector using EDI technology (see part 1 above).

EDIFACT is the traditional standard for EDI. But more and more XML is taking the scene. What does this mean for the collection and dissemination of statistics in the ESS?

EDI and EDIFACT

EDIFACT was the pre-Web standard for EDI. The typical communication media were dedicated lines or Value Added Networks (VANs). The explosion of the Internet does not automatically mean the end of EDIFACT: EDIFACT messages can very well be sent over the Internet, as e-mail attachment for example, or via file transfer.

EDIFACT became a UN standard in 1987. From the beginning, the ESS participated actively in the definition of EDIFACT messages. The relevant EDIFACT body is EBES, the European Board for EDI Standardisation. EBES is organised in different expert groups according to the different domains specifying EDIFACT messages. EBES Expert Group 6 (EEG6) is the expert group for statistical purposes.

For specific statistical domains subsets of the relevant EDIFACT messages developed by other competent bodies were adopted. Foreign trade statistics, for example, use subsets of customs messages specified by the customs expert group of EBES (EEG3). However, the statistical community, i.e. EEG6, also developed their own generic EDIFACT messages: GESMES for multi-dimensional data or chronological series, CLASET for the exchange of classifications and RDRMES for raw data reporting. Each of these messages is maintained by a specific working group within EEG6.

In some domains, EDIFACT was widely accepted. The European System of Central Banks (ESCB) adopted GESMES as their data transmission format (in fact a specific dialect of GESMES named GESMES/CB). Data exchange between Eurostat and its partners (NSIs, ECB and others) is also based on GESMES. The EDIFACT message for Intrastat reporting (CUSDEC/INSTAT) is used by over 50,000 enterprises; several million copies of this message are sent every year.

In other domains, the acceptance of EDIFACT was weaker. This is particularly the case when small and medium sized enterprises (SME) are involved. They are reluctant to introduce EDIFACT solutions for several reasons: high complexity, high costs, little flexibility, to name only the most important. These problems may be overcome by a new approach, especially tailored to the needs of the Web: XML.

XML – the Web approach for electronic data exchange

There are three languages that are often mixed up: SGML, HTML and XML. XML is the youngest in this row. Why was XML developed? Why wasn't SGML taken as the Web language? And why not stick to HTML?

SGML is a general document description language. It was conceived before the Web came up (adopted by ISO in 1986). SGML allows the definition of different document types, in other words, of different languages describing specific types of documents. In this sense SGML is a meta-language. And SGML is quite complex – too complex for the Web (making implementation expensive and applications slow).

HTML was developed as the document description language of the Web (around 1990). HTML is one specific SGML application i.e. it defines one fixed type of document. The problem with HTML is: it is not extensible, there will always be applications that cannot be based on HTML.

XML was conceived to overcome the problems linked to SGML (overly complex) and HTML (inflexible). XML is a subset of SGML, designed to enable the use of SGML functionality on the Web. Like SGML (and unlike HTML) it is a meta-language, allowing you to specify your own languages for your specific purposes.

XML 1.0 has been adopted by the World Wide Web Consortium (W3C) in 1998. XML is well suited for WebEDI i.e. for EDI over the Web. XML is supported by the major players in software industry (Microsoft, IBM, Sun, Oracle, …) and by international administrations (OECD, IMF, ECB, EU, …). A further advantage of XML is: it makes business-to-consumer (B2C) communication possible; this was not the case with EDIFACT – which individual or household would have installed an EDIFACT solution at home, expensive and complex as it is (as long as it is not offered for free)? XML however is accessible using a normal browser. XML forms, made available by businesses on their Web site, can be used by everybody – no extra costs arise for the consumer.

So XML is the ideal candidate for the future of EDI. But XML is only the basis. To do WebEDI you need a bit more.

ebXML – electronic business based on XML

XML is only the syntax. Semantics are defined separately. This is where ebXML comes in.

ebXML (electronic Business XML) is a joint venture of OASIS i.e. the software industry (including Microsoft, IBM, SUN, SAP, …) and UN/CEFACT, the international business standards body responsible for EDIFACT. ebXML has defined an XML framework (core components, business processes, registry services, …) for electronic business. On 14 May 2001, the approval of the ebXML specifications was announced to the press:

"Geneva, Switzerland and Boston, MA, USA; 14 May 2001 – UN/CEFACT and OASIS today announced that participants from around the world approved ebXML specifications at a meeting in Vienna, Austria on 11 May 2001. ebXML, which began as an 18-month initiative sponsored by UN/CEFACT and OASIS, is a modular suite of specifications that enables enterprises of any size and in any geographical location to conduct business over the Internet. Using ebXML, companies now have a standard method to exchange business messages, conduct trading relationships, communicate data in common terms and define and register business processes." Source: http://www.ebxml.org.

ebXML is supported by EBES. It is declared Eurostat strategy to develop and introduce ebXML compliant versions of the statistical messages used in the ESS. The basis for these developments will be UML data models (UML = Unified Modelling Language). Eurostat's data transmission and dissemination tools (like STADIUM or NewCronos) will support XML. However, the existing EDIFACT solutions will be maintained as long as required.

Other semantic standards based on XML

ebXML is not the only XML-based semantic standard. Basically each field of activity can have its own specific set of standards built on XML. Some of those standards could be used for statistical purposes: compliant databases could be interrogated automatically by statistical questionnaires, for example.

The eXtensible Business Reporting Language (XBRL) is an XML-based specification for the preparation and exchange of financial reports and data. XBRL is driven by the accounting profession, in particular the American Institute of Certified Public Accountants. XBRL uses accepted financial reporting standards and practices to exchange financial statements across all software and technologies, including the Internet. More under http://www.xbrl.org.

As mentioned above, the IQML project is using the Common Warehouse Metamodel. CWM is the result of a request for proposal issued by OMG in 1998. Partners in the project include IBM, Unisys, Oracle and others. "The purpose of OMG's Common Warehouse Metadata Initiative (CWMI) is to enable easy interchange of metadata between data warehousing tools and metadata repositories in distributed heterogeneous environments." Source: http://www.cwmforum.org.

The XForms project was launched by W3C. They claim that XForms is the XML language for the next generation of Web forms. "XForms is W3C's name for a specification of Web forms that can be used with a wide variety of platforms including desktop computers, hand helds, information appliances, and even paper." Source: http://www.w3.org/MarkUp/Forms.

There are many more XML applications and industry initiatives; a good source of information is http://xml.coverpages.org. While ebXML has been chosen as the semantic XML standard for the ESS, the other standards should not be lost sight of. Under special circumstances they may be the better solution, and you should at least learn what they do better.

Part 3: Dissemination

Electronic dissemination of statistical results is getting more and more important. Users want immediate access to the newest results, not only at national but also at international level, in a language they can understand; usually, English is the language used to reach an international audience. Permanent availability of statistics (24 hours a day, 7 days a week), continual user support and free access to statistics are other user requirements. Very important, and sometimes neglected, is the usability of the Web services offered.

A European benchmarking

A benchmarking study on electronic dissemination carried out by Statistics Denmark in 2000, comparing 7 European countries, shows certain trends [Statistics Denmark 2001]:

- the percentage of households with Internet access visiting the NSI's Web site is quite high, in some countries over 50%.

- NSIs tend to make all publications available on the Web for free, mostly in PDF format; in some countries the Web is the first priority dissemination medium for publications. A problem with PDF is that it is not easily searchable by software tools. Interestingly the sale of hard copies seems to increase when a publication is available for free on the Web, but there is no scientific evidence yet for this.

- databases become accessible over the Web; examples are StatLine in the Netherlands or StatBase in the UK.

- free database access is offered more and more; nevertheless, in most cases users have to register before gaining access.

- a shift towards XML as an output data format is expected in the future; it is, however, not widely used today, says the study.

- user satisfaction is linked to certain criteria, including: the data needed is available (make as much as possible available on the Web); the user is able to find the data (good search facilities; English translation available); and the user is able to understand the data (appropriate documentation for different types of users).

Usability testing

Of course, users expect accurate, timely and reliable statistical data. This is not new. With electronic publication on the Web, however, the definition of these qualities becomes more stringent [Levi 2001]:

- with many thousands of anonymous users accessing data on the Web, corrections will not reach them all; this means data have to be really accurate on the Web.

- data dissemination over the Web is instantaneous, users do not accept any delay; the interval between publication and the first access is typically a few seconds only.

- user expectation of availability of service is raised; Web site overload or down times will not be accepted; global access requires continual availability all around the clock.

With the Web an additional challenge for statistical agencies has emerged: "presenting complex data in a form that can meet the differing needs of a highly diverse population." This problem of usability is a non-trivial task. Usability engineering and usability testing offer methods to measure and improve the design of Web sites.

This problem should not be underestimated. The way a statistical agency is perceived and accepted by its users will depend more and more on the usability and quality of what it offers on the Web.

Researching the future

Statistical databases on the Web may not be the ultimate solution users want. Users may bok for information from different sites, presented in a harmonised fashion. The EU research project MISSION (Multi-agent Integration of Shared Statistical Information Over the [inter]Net) takes this road.

MISSION (http://www.epros.ed.ac.uk/mission/) aims to give access to distributed heterogeneous data sources with minimum effort. The project is based on agent technology and XML-based description of metadata. A software suite will be developed forming a data dissemination network, or better: a data access network. The elements of this network are:

- the Client: this is the user's access point. The Client is the MISSION network user interface; it is a Web application connecting the user to a Library.

- the Library: Libraries are repositories holding statistical metadata. Metadata support the access to statistical data (access metadata), the processing of data for analytical purposes (methodological metadata) and supply background information (contextual metadata). As Libraries communicate with each other, the user has virtual access to all Libraries. Libraries access Data servers to search statistical data.

- the Data server: Data servers can be seen as gateways to data stores like the NSI's databases or data warehouses.

- Agents: while Clients, Libraries and Data servers are the static units of the MISSION network, agents are the active operators. Agents navigate the Web to locate and access the appropriate static units, where they invoke the required operations to satisfy the user's request.

Suppliers of statistical data will be able to subscribe to the MISSION network. MISSION offers an interface to their existing data via a Data server. The supplier will, however, retain control over all aspects of access to their data.

A user will be able to make declarative requests via a Client. No expert know-how will be required. Results will be harmonised, manipulation methods well documented.

The innovative aspect of MISSION is the use of agent technology. The system will permit easy access to data available on different technical platforms in heterogeneous data storage systems.

Conclusion

The Internet revolution has had a profound impact on both domains, statistical data collection and dissemination. This has changed the ESS and will continue to do so. The end point of this development is not yet in sight: new developments show up every day – the Web is barely 10 years old, the XML recommendation 3 years and ebXML not yet 1 year. What will come next?

The major changes we have seen so far include:

– intelligent electronic questionnaires and Web forms.

– the rise of XML and ebXML.

– Web publishing and agent technology.

These changes bring benefits to all parties involved in the statistical process – data providers (lower burden), statistical agencies (reduced production costs) and users (improved timeliness and quality). But they also mean challenges – this, however, mainly for the statistical authorities: they have to adapt their systems on a permanent basis.

Glossary

B2B	Business-to-Business
B2C	Business-to-Consumer
CEFACT	UN Centre for the Facilitation of Practices and Procedures for Administration, Commerce and Transport
CLASET	Classification Exchange and Transfer message
CN	Combined Nomenclature
CSAQ	Computerised Self-Administered Questionnaires
CUSDEC	Customs Declaration message
CWM	Common Warehouse Metamodel
CWMI	Common Warehouse Metadata Initiative
EBES	European Board for EDI Standardisation
ebXML	electronic Business XML
ECB	European Central Bank
EDI	Electronic Data Interchange
EDIFACT	Electronic Data Interchange for Administration, Commerce and Transport
EDISENT	EDI between Statistics and Enterprises
EDR	Electronic Data Reporting
EEG6	EBES Expert Group 6 – Statistics
ESCB	European System of Central Banks
ESS	European Statistical System
EU	European Union
EUROSTAT	Statistical Office of the European Communities
GESMES	Generic Statistical Message
HTML	HyperText Markup Language
ICT	Information and Communication Technology
IDEP/CN8	Intrastat Data Entry Package with the CN at 8 digit level
IMF	International Monetary Fund

INSTAT	CUSDEC/INSTAT – Intrastat subset of the Customs Declaration message
Intrastat	Statistical system relating to the trading of goods between EU Member States
IQML	Intelligent Questionnaire Mark-up Language
ISO	International Organisation for Standardisation
MISSION	Multi-agent Integration of Shared Statistical Information Over the (inter)Net
NewCronos	Eurostat time series reference database
NSI	National Statistical Institute
OASIS	Organisation for the Advancement of Structured Information Standards
OECD	Organisation for Economic Co-operation and Development
OMG	Object Management Group
PDF	Portable Document Format
RDRMES	Raw Data Reporting Message
SGML	Standard Generalised Mark-up Language
SME	Small and Medium sized Enterprises
STADIUM	Software used for the data transmission between Eurostat and its partners
TELER	Telematics for Enterprise Reporting
UML	Unified Modelling Language
UN	United Nations
VAN	Value Added Network
W3C	World Wide Web Consortium
WebEDI	EDI over the Web
WWW	World Wide Web – also known as 'the Web'
XBRL	eXtensible Business Reporting Language
XForms	W3C initiative on Web forms
XML	eXtensible Mark-up Language

References

- TELER Final Report. Philippe Caille (CESIA), Louis-Aimé de Fouquières (CESIA) and Hans Stol (Statistics Netherlands). May 1999.
- Web-forms for Intrastat. Eurostat, March 2000.
- Electronic Collection of Raw Data (eCoRD) – a European Perspective. Uwe Kunzler, Eurostat. EDR workshop, Hull, Canada, September 2000.
- Electronic Dissemination – an International Benchmarking 2000. Statistics Denmark, February 2001.
- Data Dissemination to a Web-Based Audience: Managing Usability Testing during the Development Cycle. Michael D. Levi, U.S. Bureau of Labor Statistics. MSIT conference, Geneva, Switzerland, February 2001.
- Intelligent Use of Metadata in the Questionnaire Design Process. Karen Brannen, CES, University of Edinburgh. NTTS / ETK conference, Crete, Greece, June 2001.
- e-Quest: a Metadata-based Software for Electronic Raw Data Collection at Statistics Austria. Wolfgang Koller and Günther Zettl, Statistics Austria. NTTS / ETK conference, Crete, Greece, June 2001.

- Electronic Data Reporting (EDR) Strategy for the European Statistical System (ESS). Uwe Kunzler, Eurostat. NTTS / ETK conference, Crete, Greece, June 2001.
- XML Messages for Foreign Trade Statistics. Jonathan Bates and Uwe Kunzler, Eurostat. NTTS / ETK conference, Crete, Greece, June 2001.
- User Interface for access to Statistical Databases, the MISSION perspective. Karel Pagrach, DESAN Marktonderzoek B.V. NTTS / ETK conference, Crete, Greece, June 2001.