UNITED
NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2005/19
24 March 2005

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-third plenary session
(Geneva, 13-15 June 2005)

eSTATISTIK.core: ELECTRONIC DATA REPORTING FROM ERP SYSTEMS

Invited paper submitted by the Federal Statistical Office, Germany∗

INTRODUCTION

1. In cooperation with renowned software producers and respondents, the German statistical offices have developed the internet-based *Common On-Line Raw Data Entry, eSTATISTIK.core*. The primary objective of the project is to provide efficient methods for generating and reporting statistical data from Enterprise Resource Planning systems (ERP systems). It is based on a systematic approach aimed at a reduction of the burden of responding companies and, at the same time, it is a big step forward in the improvement of data collection procedures. eSTATISTIK.core is also integrated in the national eGovernment initiative “Deutschland Online” and has been submitted for national standardization.

2. eSTATISTIK.core consists of several infrastructure and software components: a single point of delivery on the web on a central data collection server – *CORE.server* –, standardized cross-survey XML[1]-based document formats, electronic survey definitions and free software, namely the software library *CORE.connect* and the stand-alone application *CORE.reporter*. eSTATISTIK.core forwards raw data from the central data collection server to the statistical office specified as the addressee and makes them available to the survey’s production process in the proper format. eSTATISTIK.core also

∗ Prepared by Michael Schäfer.

checks incoming data for validity and generates XML-based protocols suitable for automated processing on the respondent side.

3. An evaluation version of eSTATISTIK.core has been available since early 2004. In the first quarter of 2005, a production version will be released and used for regular application in selected surveys.

MOTIVATION AND BEGINNING

4. Reducing the burden of respondents is high up on the political agenda. Unfortunately, this multi-faceted task does not lend itself to simple solutions. From the subject-matter perspective, standardizing and simplifying questionnaires and terminology, cutting down on questions and frequency, and resorting to already available data are – sometimes – possible approaches. When it comes to actual data collection, the task is narrowed down to assisting respondents in providing data. Here, many statistical offices have developed internet-based solutions that facilitate accessing and filling in questionnaires and uploading raw data files. Such solutions often go as far as providing a manual interface for data delivery, which is sufficient for individuals and smaller businesses in general, reasonably-sized questionnaires and modest data volumes.

5. In business surveys, a relatively small number of sources supply very large volumes of raw data. Typically, these sources are either big companies responding on their own behalf or service providers reporting as a third party in the name of customers. In total, service providers represent a large number of small and medium businesses.

6. Many large companies and service providers use sophisticated ERP systems, mostly from one of the few major vendors, or other IT systems for managing business data. In the past, though not in all cases, functions for generating statistical messages have been implemented by ERP software vendors or the businesses themselves. However, many of these implementations come with a number of disadvantages: they are survey-specific and require maintenance even in the case of smaller changes to the survey; the variety of survey data interfaces makes software re-use difficult, and they allow for little, if any, automation of the reporting procedure. In addition, the federal construction of the German statistical system often forces businesses to report – for a given survey – separately to a number of statistical offices. This leaves plenty of room for improvement, and accordingly, the German industry, represented by its corporate bodies, has proposed that modern, unified, internet-based reporting procedures be developed that will make automatic compilation and transmission of statistical reports from ERP systems possible.

7. At that time, the German statistical offices had already taken first steps towards modernizing data collection, which included the development of DatML/RAW, an XML-based cross-survey document type for raw data messages, and concepts for its use as a unified raw data interface on which automated generic server-side data collection procedures could be built. The proposal of the German industry offered a good occasion to begin a discussion with the user community about those concepts, to further advance and enhance them, and to put them into practice.

8. In early 2003, a working group was established with the objective to discuss

improvements of data collection procedures involving businesses and organisations, and to prepare a pilot for selected wage statistics. This working group represents the German statistical offices and partners from the industry, mainly through the *AWV - Arbeitsgemeinschaft für wirtschaftliche Verwaltung* (Working Party for Economical Administration). The AWV works towards improving the relationship between the economy and the public service. It counts among its members respondents, service providers and a total of more than 70 software producers, many of which are of great importance and influence, like Lufthansa, Datev, SAP, Oracle and UBM, to name only a few.

9. eSTATISTIK.core is not an isolated project. As a central standardization effort, it is embedded in the common strategic program of the German statistical offices, called *Masterplan*, which has two main objectives: making the German statistical system more effective and lessening the burden of respondents. The Masterplan improves cooperation and coordinates modernization activities, many of which relate to the term *eSTATISTIK.* eSTATISTIK is a strategic initiative aimed at bringing statistical services online and modernizing the statistical data collection, production and dissemination systems. Every measure in the context of eSTATISTIK relies on a thorough analysis of the relevant process and the use of standards in its implementation.

OBJECTIVES

10. Although the working group’s principal purpose was to find ways for easing the burden of respondents, it soon became clear that the groundwork for any solution had to be laid inside the statistical system and that it would have to include a modernization of back-end data collection procedures. Therefore, the objectives that have been identified by the working group take the needs of both the respondents and the statistical system into account.

11. The envisioned solution is to produce substantial improvements in the following areas: creation and client-side validation of raw data messages in general, and standardization and unification of formats as a precondition for reducing implementation and maintenance expenditures; electronic transmission of raw data messages to the statistical offices, and of acknowledgements and protocols to the respondents or senders of the messages; customer-oriented services.

12. The response burden is to be reduced to the greatest possible extent. This includes the elimination of manual work through automation of reporting procedures, the reduction of costs by means of software standardization and re-use, and improvements of the data collection infrastructure.

13. The automation and unification of data collection procedures will increase the efficiency of the statistical system. The statistical offices will further profit from receiving pre-checked, high-quality data, and improve timeliness and accuracy of their products.

14. It is expected that the cooperation between individual companies, economic organizations and the statistical offices will ensure that software producers and respondents accept and use the new data collection procedures, and that this partnership will raise the reputation of the statistical offices as modern, customer-oriented service and information providers.

REQUIREMENTS AND LIMITATIONS

15. With modern Internet technologies, data can be transmitted in a cost-effective, fast and secure way. Therefore, eSTATISTIK.core is Internet-based. Responding businesses must have access to the Internet, which is commonly the case nowadays. For transferring large documents, a broadband connection is of advantage.

16. Respondents will be able to pack any number of reports, in any combination of survey and addressee, into a single file, and send it to a single point of delivery. Also, respondents will have a way to request acknowledgements and access and process electronic protocols to track reporting activities and handle errors that occur in the process.

17. In order to keep implementation and maintenance costs at bay – especially in the medium and long term – the necessity for survey-specific implementation work should be minimized. This requires that as many generic software components be developed and used as possible.

18. The statistical offices provide the metadata that controls eSTATISTIK.core. They must guarantee the metadata's accuracy and validity and their ability to sustain the data collection infrastructure.

19. The robustness of the data and metadata interfaces – or document formats – and their suitability for generic processing across surveys is of central importance. Version control is indispensable for dealing with compatibility issues, and platform-neutrality an absolute must, given the diversity of hardware and software platforms that ERP systems and statistical production systems are run on. The Extensible Markup Language (XML) is the state-of-the-art technology for developing document types that meet these requirements. A wealth of (often free) software is available for creating XML-based applications, often implemented in the Java language that offers a particularly good support of XML processing and is portable to many platforms. Consequently, all documents exchanged between respondents and statistical offices will be XML documents.

20. It is incumbent on the statistical offices to pave most of the way for the improvement of data collection procedures and services. However, any of their contributions can only go up to the point where the integration into a respondent's specific software and hardware environment begins, but they can make that task easier.

21. eSTATISTIK.core has a potential to produce significant savings. It will probably turn out to be economical in most cases, but whether is or not for an individual respondent depends on many factors outside the scope of this solution and is difficult to predict. However, the eSTATISTIK.core project is central to the modernization of the statistical system and it is important to measure its effects. Accordingly, a study is currently conducted

that will provide a clear picture of the impact of eSTATISTIK.core on the response burden.

OUTLINE OF THE SOLUTUION

22. Bearing the above-mentioned requirements and limitations in mind, the partners involved in the project agreed that eSTATISTIK.core should centre on improvements of data and software interfaces and should have the following key features:

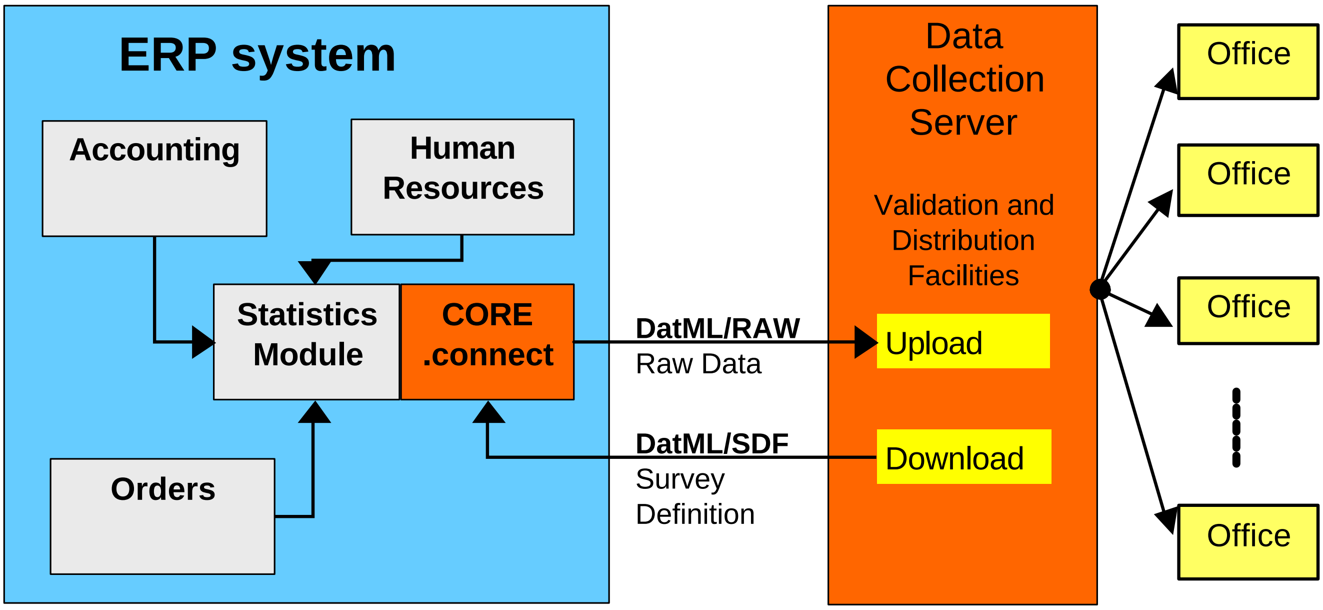
- XML-based, generic document types for raw data, survey definitions, validation protocols and acknowledgements; for raw data, the document type DatML/RAW would be used, and document types for survey definitions – DatML/SDF – and acknowledgements and protocols – DatML/RES – would be developed;
- a software library, CORE.connect, implemented in a variety of common programming languages, for client-side implementation with the following functionalities: validation and upload of raw data messages, and download of survey definitions and validation protocols;
- CORE.reporter, a stand-alone application built on CORE.connect, with a graphical user interface and facilities for managing reports and for mapping data from existing sources such as flat files onto survey data models;
- so-called *statistics modules*, implemented by the vendors of ERP systems, that extract raw data from an ERP system, generate a DatML/RAW document and upload it with CORE.connect to the data collection server. Such modules could be implemented in a generic manner, using DatML/SDF survey definitions and flexible mapping mechanisms instead of hard-coding the linkage between raw data and variables;
- a central internet server for uploading raw data messages to a single internet address and for downloading survey definitions and protocols;
- suitability for fully automated operation.

23. In its first version, eSTATISTIK.core does not support the generation of DatML/RAW documents. Applications construct a DatML/RAW document and pass it to CORE.connect for validation and upload to the data collection server.

24. Beyond the purely technological approach described above, it has also been agreed that the terminologies used in statistics and business will be harmonized to achieve a higher degree of conceptual congruence and to minimize the necessity of adapting business data to statistical concepts.

ARCHITECTURE OF eSTATISTIK.core

25. Overview:



26. The overview graphic depicts the architectural concept of eSTATISTIK.core: a) within an ERP system, a statistics module retrieves raw data from subsystems such as Human Resources and Accounting and compiles them into a DatML/RAW document; b) the DatML/RAW document is validated against an XML schema and against a DatML/SDF survey definition, and – if valid – c) uploaded to the central data collection server, where d) it is decompiled into single raw data messages which are then e) forwarded to the collecting statistical office.

27. Client-side processing depends in part on how a statistics module is implemented. Basically, a statistics module downloads the survey's current DatML/SDF survey definition and uses it to retrieve the requested raw data. In the current version of eSTATISTIK.core, the statistics module must generate the entire DatML/RAW document, which is then passed to CORE.connect for validation and upload to the data collection server.

28. On the data collection server, ingoing documents are submitted to a series of processing steps, each implemented in a separate module.

29. *Selector* checks the type of a document. Documents of a valid type other than DatML/RAW are handled the "traditional way", that is, they are passed to the data collection subsystem that processes non-XML documents. If a document is a DatML/RAW document, it is further validated against the XML schema.

30. *Inspector* validates each raw data report contained in a DatML/RAW document against the matching survey definition loaded from a DatML/SDF document.
31. *RawFilter* decompiles a DatML/RAW document into one or more DatML/RAW documents; the number of result documents depends primarily on the number of raw data

reports contained in the input document and the configuration of RawFilter. For instance, RawFilter can be configured so that each result document contains a single raw data report, or all reports from the input document with a specific collecting office or survey.

32. *Raw2Flat* converts DatML/RAW documents into flat files of a variety of formats (EBCDIC, ASCII, CSV). This module is necessary because most surveys still use flat file formats as input to their production process. When converting a DatML/RAW document, Raw2Flat loads a data set definition file that contains a description of the data set structure plus instructions on how to create the data set from a selected raw data report.

33. *Forwarder* is responsible for transferring documents generated by RawFilter and/or Raw2Flat to the collecting statistical office. It uses either FTP/SFTP or DVE, an application developed by the statistical offices for exchanging data sets in a standardized way.

34. *Protocol* creates DatML/RES validation protocols. Modules that contribute to the validation protocol (currently: Selector and Inspector) use the Protocol API[2] to store protocol data at designated nodes in the file system. At intervals, the Protocol server scans those file system nodes; if it encounters a node that is marked complete, a DatML/RES document is constructed from the node's contents and then the node is deleted. In the future, the Protocol API may use a database to store temporary protocol data instead of the file system.

XML DOCUMENT TYPES

35. DatML/RAW is a generic – that is, survey-independent – document type for raw data messages. It was first released in 2001 and is used in more than a dozen surveys. A DatML/RAW document contains at least one *message* that contains at least one *report*. Its flexible structure can accommodate any number of raw data reports in any combination of survey, reference period, respondent and collecting office. Within a message, metadata can be shared among reports. DatML/RAW stores survey-specific raw data in a generic structure of elements representing *records*, *variables* and *variable groups*. This generic structure varies from survey to survey alone in the number and nesting of these elements and in the names attached to them.

36. DatML/RES is a document type for acknowledgements and validation protocols. A validation protocol contains information about the input document and describes the validation parameters and the validation results at the document, message and report level, including the type and position of errors, and the affected variable, if any.

37. A DatML/SDF document holds the key metadata needed to apply eSTATISTIK.core to a specific survey. It describes how a survey is identified, its general characteristics such as the reference period and the periodicity, its data model – that is, variables and variable groups, and their characteristics and dependencies – and how the data model maps onto the logical structure of a raw data message.

METADATA MANAGEMENT

38. For running eSTATISTIK.core effectively, a considerable amount of metadata are necessary, growing with each survey added, and with each change made to existing metadata

resources. Furthermore, invalid metadata might compromise the whole system. Once in a productive state, such a system cannot be run on metadata of dubious quality. Therefore, a variety of tools are used to create the metadata for eSTATISTIK.core, and a new metadata management system is being built up.

39. *STATSPEZ*[3] is a metadata-based tool for specifying and creating tabulation programs. One of its components is the *data set designer*, which provides the data set definitions, including the mapping of variables, which control the converter Raw2Flat.

40. The *data edit designer* ("PL-Editor") has primarily been developed for specifying data edits, but it is also the tool with which variables definitions are created.

41. The *survey definition editor* ("SDF-Editor") has been designed for creating DatML/SDF survey definition documents. It uses variable definitions from the data edit designer.

42. In the context of eSTATISTIK.core, survey definitions and data set definitions (for converting DatML/RAW documents into flat files) are the primary metadata resources. In the context of a complete survey, however, many more metadata resources have to be dealt with. For this reason, a new metadata management system is being built up to make metadata management feasible through the whole life cycle of a survey, and beyond. This system uses three conceptual elements for organizing metadata: statistics, survey and resource. When an instance of any of these elements is released, it is assigned a unique identifier. This identifier manages resources in the context of a survey, and includes version control.

OUTLOOK

43. A future version of the library CORE.connect will offer programming interfaces that eliminate the need to deal with document format issues. Instead of creating a DatML/RAW document, applications will be able to pass data only and rely on CORE.connect to construct a valid document.

44. In addition to the HTTPS protocol, CORE.connect will support OSCI[4], a protocol which supports public/private key infrastructures, and which has been standardized for eGovernment applications.

45. DatML/RES validation protocols will indicate the location of an error in a DatML/RAW document with XPath[5], a language for constructing semantic pointers to locations in XML documents. With XPath, erroneous data can be retrieved directly from the DatML/RAW document and, for instance, presented for correction, together with information taken from variable definitions in the associated DatML/SDF document.

46. To ensure the success of eSTATISTIK.core, a series of marketing and promotion activities will be conducted by the statistical offices, the AVW, businesses involved in the pilots, and software producers.

CONCLUSION

47. eSTATISTIK.core is a new approach to improving data collection from businesses. Its foundations are standardized, XML-based raw data interfaces and metadata objects. They make it possible to set up generic, automated reporting and collection procedures with a great potential to take away a good part of the respondents' burden and at the same time to improve the efficiency of the statistical system and the quality and timeliness of its products.

48. However, standardized raw data and metadata interfaces alone are not enough to ensure the success of eSTATISTIK.core. On the part of the statistical offices, good quality metadata and a reliable metadata management are indispensable for running the system productively, and it would be bound to fail without contribution and support from the targeted user community.

[1] Extensible Markup Language; see http://www.w3.org/XML
[2] Application Programming Interface.
[3] Statistische Tabellenspezifikation; see http://www.statspez.de
[4] Online Services Computer Interface;see http://www.osci.de
[5] XML Path Language; see http://www.w3.org/TR/xpath

* * * * *