WP.32
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (v): Changing organizational cultures

A Primer in ESS Data Editing as a Dynamic System: TRIS

Invited Paper

Prepared by Laur Ivan, Eurostat, European Commission

Abstract

The European Statistical System (ESS) is a continuously evolving driver, where data quality is of central focus. Data editing, an essential component in quality measurement and control, is continuously advancing from methodology and technology perspectives. This paper presents methodological and architectural information approach in data editing through a state-of-the-art informatics framework for collecting, managing and disseminating Community transport statistics (TRIS). We follow its architecture, defining and detailing its data editing components. We outline some of the challenges in data editing with a view of increasing ROI for all ESS members.

Keywords: Data editing, statistical process, IT architecture, ESS

I. Introduction

1. Eurostat's mission is to provide the EU with high-quality statistics at European level. Therefore, data validation is of central focus. Currently, data validation in Eurostat is highly heterogeneous across statistical domains and is accomplished using tools developed independently for each statistical process (the so-called "stovepipes").

2. Improving and harmonising data validation across statistical domains in Eurostat is an ongoing process. It has been reinforced by the adoption of the Commission Communication COM (2009) 404 of 10.8.2009 "Towards an integrated method for the production of EU statistics", and the following ESS Joint Strategy adopted in May 2010. It calls for both horizontal (cross domains) and vertical (within the ESS) integration and therefore the availability of harmonised tools to produce and disseminate statistics in the ESS. It also advocates in favour of migrating from the stovepipe approach to an integrated approach.

3. **TRIS** (TRansport Information System) is an Information System framework integrating the processing of the different transport modes Road Freight, Rail, Inland Waterways, Maritime and Aviation statistics (based on Council and Parliament Regulations), and specific applications like Regional transport statistics (an on-line questionnaire) and the on-line Common Questionnaire in transport statistics together with UNECE and OECD. The process workflow concentrates on the reception, validation, processing and compiling indicators in form of tables related to the modal transport statistics.

4. The main purpose of the TRIS project is the IT integration of processing for sub-domains which vary considerably in nature (data provider, type of data, confidentiality, statistical units, volume, data format and frequency).

5. In this paper we will present and discuss an approach in data validation through a state-of-the-art informatics framework

II. Validation in TRIS

A. Defining Data Validation

6. From a methodological point of view, the three main components of data validation are [DiMeglio2009]:

- Data editing – checks designed to identify erroneous entries (e.g. missing, inconsistent),
- Missing data and imputation – Analysis of imputation and reweighting methods designed to correct missing data caused by non-response or unresolved entries and
- Advanced validation – Advanced statistical methods (e.g. outlier detection) used to further improve data quality.

7. For Eurostat, validation is mainly understood as error detection (data editing). Very few imputations are carried out autonomously by Eurostat. We further identify 3 levels for data validation:

- The 1st level validation is the Record and File Level Validation. The received dataset is checked to be correct at the record level (the structure of the dataset, code lists, etc.) as well as the file level (e.g. sums are checked, dependencies between records are checked);
- The 2nd level validation is the Intra-Data Provider Data Validation. The received dataset is checked against other data submitted by the same data provider, and consists e.g. in checks between tables and time series checks;
- The 3rd level of validation is the Inter-Data Provider data validation. The received dataset is checked against data submitted by other data providers (mirrors checks).

8. Although from a technical point of view all levels listed above can be performed at either point (data provider or Eurostat), data required may or may not be available. 1st level occurs at the end of the data collection and is typically related to basic data. Member States or other entities collecting the data are responsible for it. The Intra-Data Provider validation concerns e.g. country-level aggregates sent to Eurostat; Member States and Eurostat perform the validation. Third level validation is usually performed by Eurostat.

B. TRIS Data Validation IT Infrastructure Components

9. The IT infrastructure for Data Validation in Eurostat is continuously evolving. The main tools used for data validation, in order of proximity to the date provider are:

- GENEDI[1] (GENeric EDI): A cross-platform EDI (Electronic Data Interchange) tool, designed to manage the tasks of statistical data transmission to Eurostat. GENEDI is distributed to MS through CIRCA and is capable of validating data (formats and codes).
- eDAMIS (electronic Data files Administration and Management Information System): is the tool implementing the Single Entry Point concept for data transmission to Eurostat.
- eVE (eDAMIS Validation Engine): A Data Validation tool leveraging the power of SDMX-ML and eDAMIS. It is capable of performing complex validation rules (e.g. conditional, evaluating mathematical expressions, code look-ups) at the file transmission phase.
- EBB (Edit Building Block): Is a generic, cross-platform data validation tool for data editing (validation, imputation, outlier detection). It is designed to provide an increased performance and flexibility [CVD Masterplan, 2007].

[1] This tool is no longer developed by Eurostat

10. Currently, data transmitted to Eurostat is, in many cases, loosely coupled to its relevant metadata. Eurostat recommends the use of SDMX (Statistical Data and Metadata eXchange) framework which unifies the two aspects in a coherent, modular architecture capable of managing the entire data flow.

III. TRIS

A. Genesis

11. Most, if not all, factors in production of statistics are continuously evolving. Technologies, collected data, business models, management processes are constantly improved towards an increased efficiency and relevance. Initially, the IT architecture of statistical domains drew independently on collections of then available tools; their linking was performed manually and in a static manner. Sub-domains of transport statistics were using tools like, eDAMIS, GENEDI, SAM, PL/SQL and SQL, LINUX scripts and ORACLE, for data transmission and processing. The process was performed manually and no metadata handler existed.

12. With the evolution of IT systems and the development of new technologies, a first TRIS system (TRIS 1) has been developed with the aim to integrate the different IT systems for processing transport statistics into a single homogeneous information system framework with the same user interface, an integrated online documentation environment and the capacity to integrate with the Eurostat centrally made available IT systems, building blocks, tools and other components while still reusing much of the initial processes and related IT code and infrastructure. In effect, TRIS 1 represented a first step towards process automation.

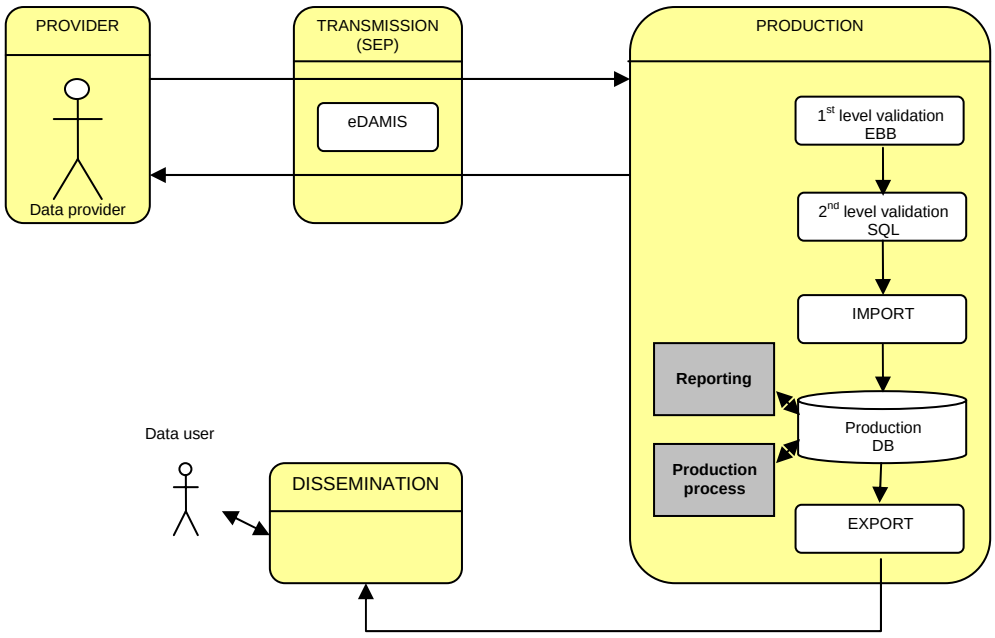

Figure 1: TRIS 1 architecture

13. TRIS 1 implements a generic workflow, common to most sub-processes:
(a) Receive data transmitted through eDAMIS (with rare exceptions),
(b) Dispatch to their respective production processes,
(c) Perform the three levels of validation,
(d) Report from each step to Member State,
(e) Perform corrections,
(f) Archive intermediate and final data validation reports,
(g) Extract/aggregate/disseminate data when declared final.

B. Drawbacks

14. From a technical perspective, the process lacks efficiency: Member States can perform first level validation outside TRIS using the GENEDI tool. Subsequently, eDAMIS is used for sending data to Eurostat. First level validation is then repeated through EBB. Second level validation is performed though PL/SQL stored procedures and third level validation is made by data manager through visual inspection of reports.

15. Besides the technology-related limitations, the process of data validation is sub-optimal. Due to its very nature, the current validation process generates a large number of information flows between Member States and Eurostat, resulting in large efforts and high resources consumption.

16. In TRIS 1, as in many EU statistics processes, data providers are expected to send clean/validated files to Eurostat which then carries out autonomous validation based on commonly agreed validation rules, returns error reports and waits for new data delivery or country specific notes on outliers. To avoid interrupting and restarting the process each time, the TRIS system allows for partial correction of data.

17. TRIS 1 does not use SDMX standards and tools for data transmission. Instead, data files are transmitted in CSV format. However, TRIS 1 can process SDMX-ML data files for early bird data providers, but converts them with Eurostat standard SDMX tools to CSV for further processing. This conversion has been implemented to give time to the many data providers to migrate to SDMX-ML.

C. TRIS 2

18. The second phase of the project, TRIS 2, is undertaking to replace as much as possible the existing IT code base with external services and tools as they become reliably, available and supported. A next step towards more automation is also planned.

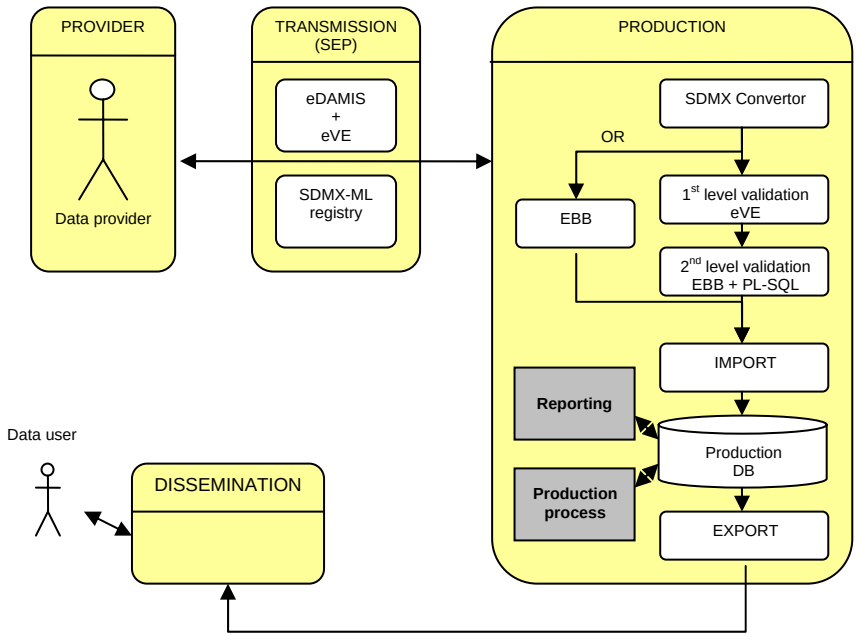

Figure 2: TRIS 2 architecture

19. For transport statistics, most of the received datasets are CSV formatted (other remaining formats like GESMES are automatically converted into CSV). Recently, the ESS has adopted a new format for statistical data interchange SDMX-ML, accompanied by a set of tools like the SDMX-ML registry, the Data Structure Wizard and the SDMX Converter, which can be used by TRIS either as external

applications, or integrated as automatically executable modules. Consequently, related actions with data providers have been initiated for migration of the existing data collections in .CSV and/or .GES to SDMX-ML.

20. TRIS 2 aims to implement SDMX-ML as a pivot format for processing, bedding on the compatibility of the available tools. A temporary limitation to using SDMX across all transport modes is triggered by the unavailability of standards for micro data processing.

21. The main driver in utilising SDMX in TRIS 2 is the compliance with Eurostat standards, resulting in a more rigorous separation of the validation levels, a 1st level validation as early as possible and as close as possible to the data managers in Member States. TRIS 2 implements SDMX components (e.g. SDMX-ML, DSD, SDMX Registry and the SDMX converter) for most transport modes (except for micro-data), following closely the standard's evolution. Thus, TRIS 2 is a pioneer and a proof of concept towards a concrete implementation of SDMX as architecture.

D. Vertical integration of validation

22. To overcome the shortcomings of having to restart validation from scratch each time data are retransmitted, TRIS 2 is building on a system of flags (which are part of eDAMIS) to differentiate between new submissions and "validation" submissions. Once a dataset has been flagged for "validation", the process ends with the last validation module (2nd level) and an error report is transmitted to the data provider. The aim of this report is to inform about any detected error giving the location and the type of the error, and to produce a summarised document containing relevant information.

23. Ideally, the **first level** validation should take place during or before transmission. In TRIS 1, Member States can perform the 1st level validation using the GENEDI tool. Drawing on SDMX-ML, TRIS 2 allows for this level to be carried out either in the Member State before transmission or at arrival in Eurostat benefiting from DSDs (Data Structure Definition) and the eDAMIS Validation Engine eVE.

24. **Second level** data validation in TRIS 1 is performed through PL/SQL stored procedures executed by TRIS. EBB is currently evaluated with a view of progressive integration in TRIS, i.e. migrate supported operations and contribute the remaining operations as feature requests.

25. With respect to integration, the second level could also benefit from taking place in the Member State, closer to dataset compilation and thus avoiding back and forth traffic involving different stakeholders. TRIS 2 aims to develop a practical solution where second level validation takes place on a central Eurostat server with automatic reporting to data providers.

26. One of the project’s medium term targets is to implement distributed validation execution, either in Eurostat or in the Member States or elsewhere where available architecture and platforms exists. Eurostat is currently developing a secure Infrastructure allowing for remote access to confidential data from data providers PCs in a secured data room that would address this challenge.

27. The **third level** data validation checks consistencies across countries within the same data collections/sub-domains. It does not involve external validation in a strict sense. This validation should normally take place later in time, at Eurostat and most probably when data from all Member States is gathered. The opportunity of performing this validation before transmission will be analysed on a case-by-case basis.

E. Other considerations

28. In addition to the components described above, and considering that a distributed architecture is not yet feasible, TRIS is analysing the possibility of reducing the number of reports by merging 1st and 2nd level reports validation.

29. The persistence of legacy formats present several challenges in data validation, e.g. the inability of using the eDAMIS Validation Engine eVE for non-SDMX-ML formats. Data collections like “Web Common Questionnaire” and processes are difficult to change due to their nature because of organisation and institutional factor like the involvement of other international organisations.

III. Conclusion and Future Work

A. Efficiency gains

30. At Eurostat, data validation is perceived as a critical step for ensuring the quality of EU statistics but also as a very costly step. Being far from data collection, the relevance of validation is however challenging and optimisation principles call for, at least partial, integration with Member States' processes. Efficiency gains can also be achieved by avoiding redundancy between data validation steps in Eurostat and Member States.

31. Striving for further integration of data validation within the ESS, two competing integration strategies are currently discussed:

(a) Perform validation in Member States (e.g. through distributing a tool like EBB) and
(b) Create a central system capable of validating data in Eurostat (e.g. via eVE/EBB) and reporting back to the Member State.

32. TRIS attempts to balance the two extremes in a coherent proposal with clear benefits for users. TRIS is seen as proof of concept and will give useful insight on the potential gains in quality and efficiency.

B. Further standardisation of Data Validation in Eurostat

33. A Vision Infrastructure Project (VIP) on Data Validation has been recently started with the explicit aim to test the feasibility and to develop a global strategy for data validation in Eurostat. The benefits from closing the gap between the collection and the validation points are viewed as essential in increasing the data quality and TRIS is viewed as an illustration of this approach.

34. Implementing the Vision will result in challenges at different levels. For example changes at technical level can manifest in re-thinking the business architecture, defining the roles of the actors (in Eurostat – in the units dealing with the processing and production of statistics as well as in the IT units - and in the ESS) and in a (re)design of Information Systems, Information Technology and Networking architectures in Eurostat.

C. Architectural challenges

35. Embracing Service-Oriented Architectures combined with a solid infrastructure for remote access to confidential data, render the location of data validation irrelevant. A complementary project for a secure infrastructure for access to confidential data, is evaluating solutions with the purpose of establishing a common infrastructure for projects with similar technological and functional requirements like Euro Group Register and the decentralised access to confidential micro data for scientific purposes.

36. One of the main challenges remains the integration of the validation service in Member States infrastructure and the sharing of common validation rules. TRIS targets in the long run to have the place of the validation execution distributed either in EUROSTAT or in the Member States or elsewhere where available architecture and platforms exists (for example by having a TRIS client installed on (secured) PCs in Member States).

IV. References

Eurostat Internal Document (2011), TRIS 2 - CAB Review, Eurostat, Luxembourg

Eurostat Internal Document (2010), TRIS 2 - Pre-analysis & Feasibility Study, Eurostat, Luxembourg

Eurostat Internal Document (2007). *CVD Masterplan.*, Eurostat, Luxembourg.

Eurostat Internal Document (2005). *Introduction to data validation.* Unit B2 (Methodology and Research), Eurostat, Luxembourg.

UNECE (2009), *Data Validation strategy in Eurostat*, Neuchâtel, Prepared by Emilio DiMeglio

SDMX Global Conference (2009), The SDMX Istat Framework. Prepared by Mauro Bianchi, Dario Camol and Laura Vignola, ISTAT, Italy

Eurostat (2005), GENEDI – Generic EDI toolbox – User Guide, European Commission

Eurostat (2007), A Single Entry Point for Data Transmission to Eurostat – European Commission, Eurostat