Distr.
GENERAL

WP.10
25 April 2012

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2012)
(Washington, DC, 21-23 May 2012)

Topic (ii): Streamlining statistical production

A corporate approach to processing microdata in Eurostat

Invited paper

Prepared by Jancsók, P. and Wirtz, C., Eurostat Luxembourg

I. Introduction

1. Eurostat, the statistical office of the European Union, works together with National Statistical Institutes (NSIs) in the European Statistical System (ESS) to provide comparable statistical information to policymakers and the public. Eurostat is also involved in providing the common methodology for data collection and preparation, it receives the data from the data providers, performs different validations, and finally compiles the European statistics.

2. The joint strategy of the ESS,[1] adopted in May 2010, aims at more efficient and integrated production methods for statistics. Sharing of IT architecture seems to be an implicit and crucial aspect of the future infrastructure of the European Statistical System and a logical step towards further integration and industrialisation of the production process.

3. Eurostat's efforts to streamline the statistical production chains for microdata processing can be divided into two main parts: internal for harmonisation and consolidation across statistical areas, and external for sharing with data providers.

4. The first part, internal harmonisation and consolidation, consists in the creation of a generic, modern production chain using state-of-the-art technologies to process microdata, covering the entire workflow from data validation to dissemination. This process was started in 2006-2007 when the Generic SAS Tool (GSAST) architecture became operational. The GSAST platform offers a uniform user interface with integrated metadata handling and clear visualisation of the workflow. In addition to the compilation of predefined statistics for standard dissemination, the environment can

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/about_eurostat/documents/ESSC20100506ENvisionfinal.pdf

also be used to answer additional or ad-hoc queries based on production data. There are now nine data collections using this tool.

5. The second part covers questions of how to cooperate more efficiently with data providers, especially in the data validation and quality monitoring steps. Currently, Eurostat is examining the possibility of providing remote access to data providers so that they can use part of the GSAST architecture to perform advanced validation steps and access the data quality monitoring functions.

II. Generic SAS Tool (GSAST) architecture

A. How microdata are processed in Eurostat using GSAST

6. The full microdata handling workflow in Eurostat can be divided into two different stages. In the first stage, which is the regular production chain, the incoming data are processed for calculating country and European level aggregated data. This production workflow has four main steps.

(a) 'Receipt of data': Data providers send the data files to Eurostat via the single entry point EDAMIS (Electronic Data file Administration and Management Information System).

(b) 'Country loop': The basic data validation is performed by Eurostat using standard tools (Pongas & Wirtz 2011). If corresponding macrodata are also transferred to Eurostat, additional quality checks are performed by comparing the transmitted macrodata with those calculated from the microdata by Eurostat. For each country, the data are validated and a deviation report is produced. In the event of errors and implausibility the relevant data providers are contacted to correct this. The 'country loop' yields clean microdata and some statistics per country. It has to be performed for all data providers.

(c) 'Aggregation': Once all or the majority of the country files have been successfully validated, statistics for groups of countries (like the EU, the euro area) are calculated, disclosure control is performed, flags are attributed, etc.

(d) 'Dissemination': Once all the relevant statistics have been computed, they are uploaded to Eurostat's reference database and disseminated via the website.[2] For some data collections anonymised microdata files are made available for researchers. Note that for the preparation of anonymised microdata files, additional special value groupings or other disclosure control steps need to be performed.

7. These four steps of processing form the regular production chain. Figure 1 shows an example of the production chain workflow.

[2] http://www.ec.europa.eu/eurostat

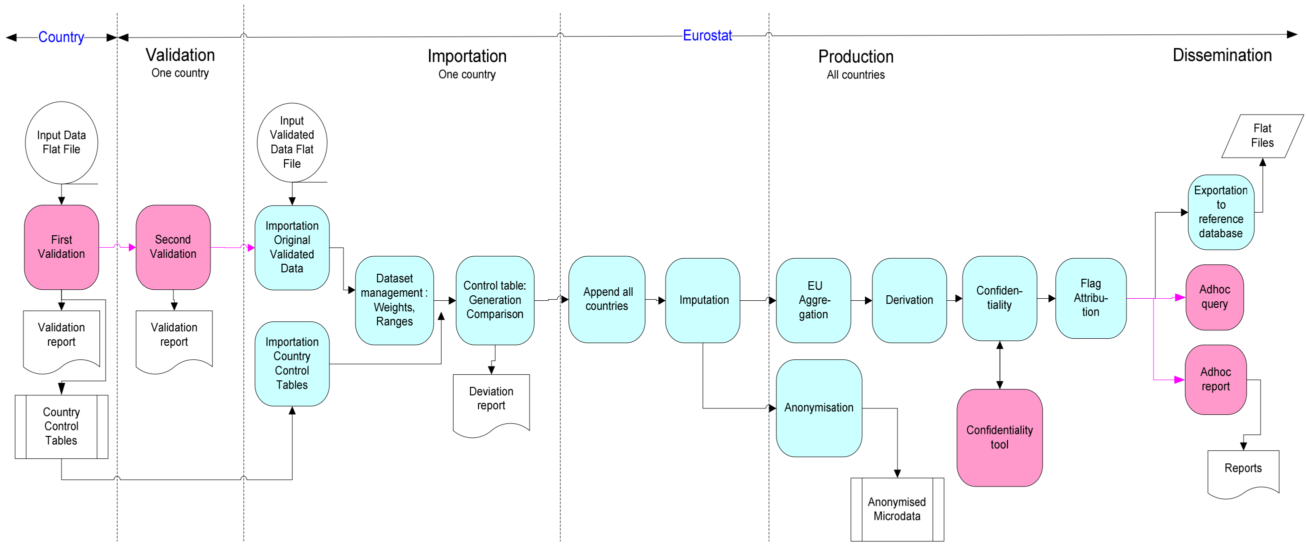

Figure 1. An example of the microdata processing workflow

8. Next to the regular production chain, which covers the standard calculation and dissemination of European statistics, there might be a second stage linked to the additional need to further analyse the data and provide statistics to answer external requests or for publications.

9. Within Eurostat the different data collections have similar workflows and this fact calls for integrated architectures and for the use of common modules in the production chain.

B. GSAST architecture, the corporate approach

10. Back in 2006, the perceived need for an integrated, centralised architecture and for a generic tool to process the increasing number of microdata collections triggered the design of GSAST. It began with a feasibility study where several microdata data collections were analysed. The outcome of this study helped to establish the functional specifications for GSAST. The main aim was to create a modular platform based on the latest technologies to help statisticians in Eurostat with the processing of microdata from the moment the data arrive at Eurostat until statistics are disseminated via the reference database. The tool should support the complete workflow of the production chains (Stage 1) while the whole environment will provide the computational base for further processing (Stage 2).

11. The specific requirements were:

(a) Use state-of-the-art technologies;
(b) Apply a centralised approach;
(c) Modular design including generic and reusable modules for all data collections;
(d) Standard user interface with a clear representation of the process workflow;
(e) Possibility to perform additional computations on the data (Stage 2);
(f) Easy to use and easy to learn by statisticians;
(g) Possibility to browse multidimensional data in an easy way;
(h) Maintenance of the tool should be easy and could preferably be performed by the statisticians.

12. To fulfil these requirements GSAST was designed and implemented as follows:

(a) The SAS Business Intelligence (SAS BI) platform (SAS Institute Inc., Cary, NC, USA) was selected as programming environment. This provided a server – client architecture and the existing experience in SAS programming could be reused. The datasets are stored in ORACLE (Oracle Corporation, Redwood City, CA, USA) databases. The disadvantage of using the SAS-based solution is that the software is linked to yearly licence fees. The confidential nature of the microdata requires a specific secured server environment.

(b) GSAST offers a corporate approach by providing a common server-based tool to the statisticians. Eurostat's IT provides centralised development and support services, while the data collection managers are able to perform fine-tuning of their applications mainly by modifying the applications metadata.

(c) The modular design is implemented in a step by step approach. If a data collection to be processed by GSAST requires an as yet inexistent functionality, then this function is developed in a generic way and is added to the pool of existing functionalities. To date there have been six waves of development and each of them has extended the available functionalities in GSAST.

(d) SAS Enterprise Guide (SAS EG) is the standard user interface. Using SAS EG the processing workflow is clearly seen on-screen as a chain of Stored Processes (represented by icons in Figure 2). Stored Processes are centrally managed SAS programs described by metadata and can have input parameters. The users are prompted for these parameters, and their possible values, data sources for dynamic prompts (e.g. tables with the available countries) are registered in the metadata server. Internal metadata tables describe the dependencies of the workflow steps and ensure that the user executes the workflow steps in the right order.

(e) Using SAS EG as a client has several advantages. The full capacity of SAS EG is available to perform additional computations. In the SAS EG environment the production data tables are available in read-only mode and all kinds of additional processing can be performed.

(f) The client provides a standard common user interface to the GSAST application, and it is easy to learn. This enables statisticians to process different microdata collections, and only the collection-specific knowledge has to be learnt while the global environment remains the same.

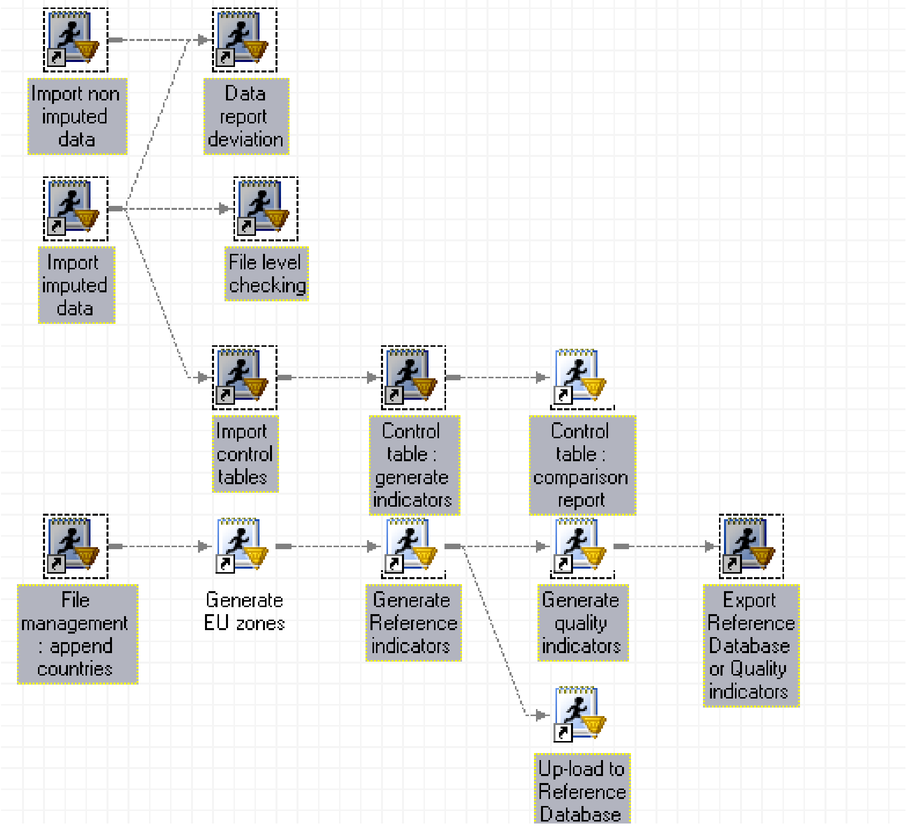

Figure 2. Example of a workflow in SAS EG

(g) Most of the microdata collections have several data dimensions. SAS EG offers a possibility to browse and visualise these data with the help of an OLAP viewer. The viewer can connect to OLAP cubes which store the multidimensional data. The cubes are accessed internally and allow easy selection of a required subset of data for a particular work. In our experience this is a very useful functionality making it easier to work with highly multidimensional data collections.

(h) To ease maintenance, GSAST was designed to extensively use metadata. All the variable parts of the production processes, including even the description of the workflow itself, are defined in metadata tables. Process parameters are either stored as metadata or the users are prompted to provide input during the process.

C. Metadata structure in GSAST

13. The working of the GSAST tool relies heavily on different metadata. For example, to be able to maintain the applications easily without the need of a programmer, all the processes, workflows, variables, indicators and aggregations are defined in different metadata tables. Metadata in GSAST can be classified by their functions to different groups.

14. *Technical metadata* related to the SAS BI server. This kind of metadata defines the detailed computational environment and is stored in the SAS Metadata server, which is the central server of the SAS BI platform. The general metadata necessary for working properly in the SAS BI environment (server definitions, user rights, etc.) belong to this group. They are defined, together with SAS libraries (collections of SAS data files), in the Metadata server and are used by the SAS EG client to access the production data and metadata.

15. *Process definition metadata* describe the processes which handle the data throughout the production workflows and can be divided into two subgroups. The first subgroup is also stored in the SAS Metadata server and includes the definitions of SAS Stored Processes with their prompts.

16. The second subgroup is stored in Oracle data tables in the tablespace of the data collection. Practically all processes are coded in a parametric way and the parameters are stored in these metadata tables. Some of the metadata tables control the workflow. One example is the ‘P_DEPENDENCIES’ table, which describes the logical sequence of the process steps and prevents the execution of a step if the previous step is still running or has failed. Other tables contain dictionaries of indicators, formats, users, or messages. Sometimes short expressions or even a piece of SAS or SQL code is stored in the metadata. This relatively heavy metadata structure allows changes in the surveys to be handled especially because it is quite common for a survey to change slightly from one edition to the next.

17. *Structural metadata* describe the datasets, variables, etc. for each data collection. In GSAST this kind of metadata is stored in Oracle tables and covers the code lists and association.

18. *Statistical metadata* are also stored in Oracle tables sometimes together with the data lines. Footnotes and flags form part of this category. In GSAST every data row has data columns providing a kind of signature. This signature is used to store the name of the user who modified/created the data row together with the date and time of modification/creation.

D. Metadata management

19. The metadata structure of the GSAST application is rather heavy and integrity is difficult to maintain if updates are manual. The maintenance for a survey mainly requires the metadata to be changed. To be able to maintain the complex metadata structures, an application was designed and developed. The main requirements of this application were:

(a) Integrated solution with the existing GSAST;
(b) Provide a user-friendly way to browse complex metadata structures;
(c) Parametric and metadata driven;
(d) Support the editing of the metadata and maintain their integrity;
(e) Support for centralised management of code lists.

20. To provide integration with the existing GSAST tool, the Metadata Editor was implemented as an add-in to Enterprise Guide. This way the same client application can be used and the existing connection to the server was used. (See Figure 3)

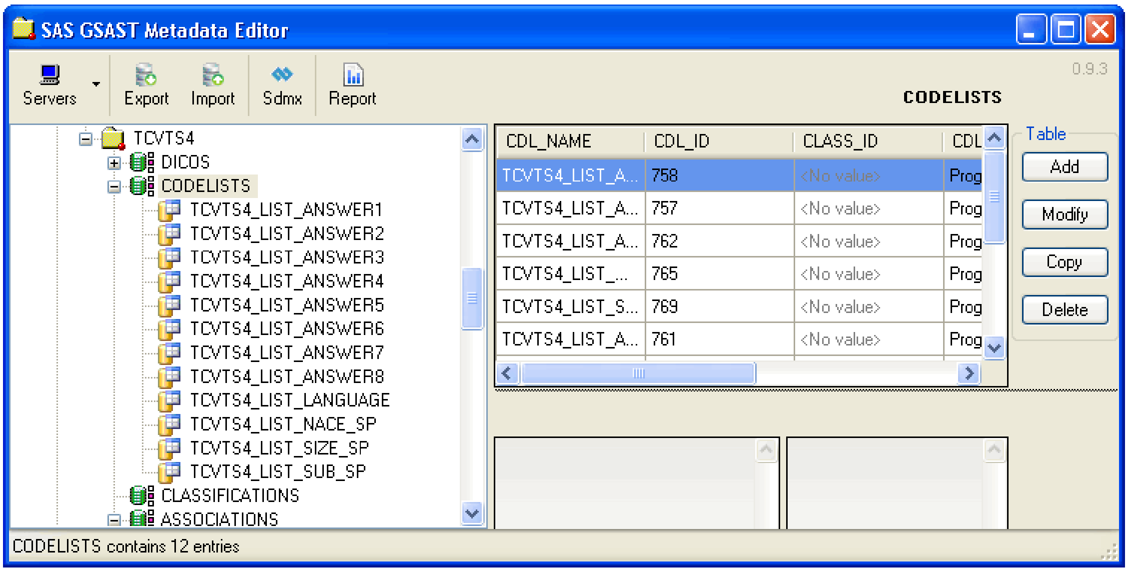

Figure 3. An example screen of the GSAST Metadata Editor

21. The simple structure of the user interface of the Metadata Editor helps new users to work with the application within a short time. This easy-to-use interface groups the metadata into different categories. This kind of grouping makes it easier to find the relevant metadata table even if the underlying structure is complex. Inside the metadata group there is a list of the relevant data files which can be viewed easily in a separate window.

22. GSAST's metadata structure is described by additional metadata tables where the relationships of the files are described. This metadata structure contains an inventory of the metadata tables by type and helps the Metadata Editor application to show the relevant tables. With the help of a flagging system it is possible to define the roles of the columns in the metadata tables. External links to other tables can also be defined.

23. Using this flagging system, validation and other rules can also be defined. A set of tools is available to guide the user through the process of changing the content of a metadata field. By means of a test bench the SAS syntax stored in the metadata field can be run and the results examined. To maintain the integrity of metadata validation, rules can also be defined in the metadata structure of the Editor. Edited metadata records can only be saved in the database if the validations are successful.

24. To maintain the centralised management of common metadata, the Metadata Editor application can connect to the SDMX Registry.[3] It is possible to search and download objects from the registry and incorporate them into the GSAST metadata tables.

E. Roles in the GSAST

25. Based on the centralised nature of GSAST the responsibilities and tasks are shared between several groups of people with different roles.

[3] http://sdmx.org/?page_id=13

(a) The first group are the Users of the GSAST. They work in the production units responsible for the data collections and their main task is to process the microdata. Their main tool is the SAS EG.

(b) The second group are the Power Users. They also work in the production units and they can maintain the production workflow by modifying the process-related metadata and developing computations belonging to Stage 2. These people typically also perform the tasks of Users. Their main tool is the SAS EG together with the GSAST Metadata Editor.

(c) The third group is the GSAST IT support. Eurostat's SAS Support group is responsible for the global development and maintenance of the GSAST. It manages the SAS servers and user accounts, deals with maintenance requests from the users and provides training on SAS EG and GSAST, etc. Its main tools are the SAS EG, GSAST Metadata Editor and the SAS Management console, as well as other IT tools.

F. Conclusion

26. The GSAST framework has evolved a lot over the years. Currently it can handle nine data collections. It is proven that the Enterprise Guide based workflows together with the stored processes can handle all the necessary processing of microdata collections. Manual maintenance of the surveys' complex metadata structure remains difficult, and requires IT and database expert knowledge. The introduction of the Metadata Editor application has simplified the process and made maintenance more user-friendly.

III. Possible extension of GSAST for usage by data providers

27. As described above, microdata processing involves close cooperation between the data provider and Eurostat. This close cooperation is mainly driven by general (e.g. Regulation No 223/2009[4]) or specific EU regulations where the details of the cooperation are strictly defined, like data quality parameters, file transfer, etc.

28. At the NSIs there are well-established workflows for collecting and processing the microdata collections which will be sent to Eurostat (Eurostat does not normally collect data directly). To ensure that coherent, accurate and comparable statistical data are received from the NSIs, thus allowing the compilation of high-quality European statistical data, Eurostat has to thoroughly validate and control the quality of the incoming data. Although the data processing and validation occur at different levels (e.g. Eurostat can compare data with its own historical database), Eurostat's workflow might overlap mainly with the final part of the processing chain of the NSIs. To rationalise the resources and optimise the cooperation inside the European Statistical System,[5] Eurostat has examined the possibility of giving NSIs access to the country loop of the GSAST processing chain.

29. This access would allow NSIs — in a self-service way — to use infrastructure of Eurostat to upload, validate and perform the (country-specific) steps of the microdata processing.

G. Current processing

30. As discussed above, the microdata processing has a country-specific component. This part of the processing needs to be executed for all countries. The data validation results might mean that the NSIs have to send updated data files, which will then be validated again, until all the issues are resolved. The full validation and quality checks could not necessarily be performed at the NSIs as it may contain steps which can only be performed at Eurostat. The country loop of the GSAST workflow has to be executed country by country and the results analysed by the person in charge of processing the data. A visual analysis of the results is often necessary for the implausibility aspects.

[4] See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:En:PDF

[5] See http://epp.eurostat.ec.europa.eu/portal/page/portal/ess_eurostat/introduction

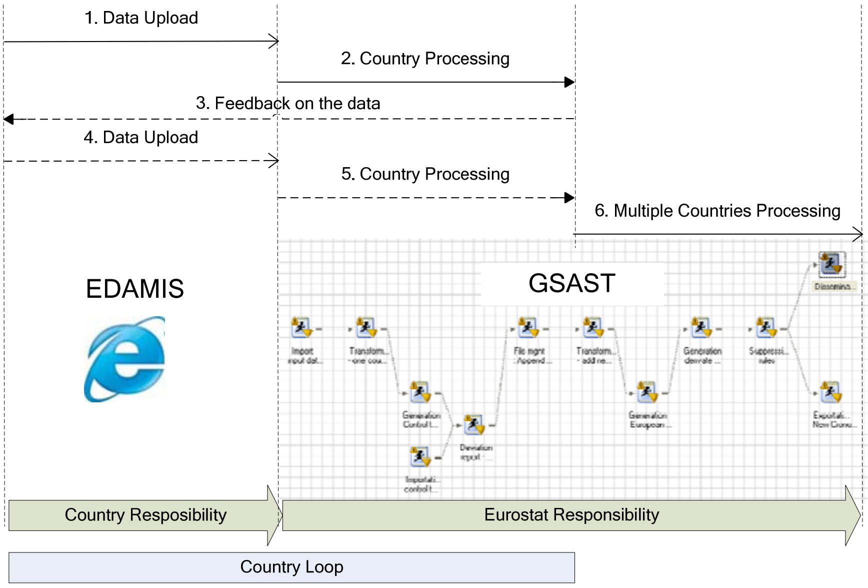

Figure 4. High-level overview of the current microdata processing

31. Figure 4 shows the current business process when using GSAST:

(a) Step 1. The survey data are initially uploaded by the NSI User using the EDAMIS system.
(b) Step 2. The uploaded data are validated and handled by Eurostat.
(c) Steps 3-4-5: Eurostat sends feedback to the data provider. The data provider sends updated data to Eurostat. These steps are only necessary when errors and implausibility are detected in the transmitted data file. They may be executed several times. This loop aims to improve data quality if Eurostat detects missing data, errors or inconsistencies in the data.
(d) Step 6: Once the data have been collected for multiple countries, Eurostat processes the data and disseminates the results.

H. Self-service approach

32. In the proposed new workflow the NSIs can have remote access to the country loop of the GSAST workflow. This way the NSIs can remotely process their data at Eurostat through the validation process. Figure 5 shows the proposed new process with the following steps:

(a) Step 1. The survey data are uploaded by the NSIs.
(b) Step 2. The uploaded data are manipulated and validated remotely by the operators in the NSIs.
(c) Steps 3-4-5: Feedback is automatically provided by the GSAST and the updated data are sent to Eurostat again. These steps are optional and can be executed several times. This loop is executed if the GSAST system detects missing data, errors or inconsistencies in the data.
(d) Step 6: Once the final data have been uploaded by the NSIs for multiple countries, Eurostat processes the data and disseminates the results.

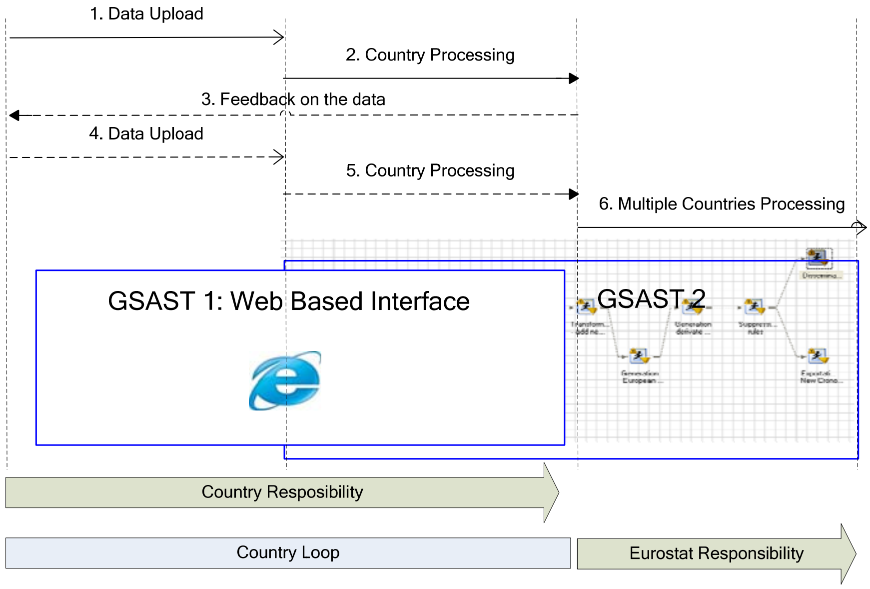

Figure 5. High-level overview of the proposed microdata processing

33. The main advantages of this system are that the NSIs can have direct feedback on their data at the time they want, and it ensures common processing rules.

I. Data security

34. To maintain data security and preserve the confidentiality of the processed data, the existing secured data transfer channels could be used. To ensure that each NSI can only process the data files belonging to their country, access rights will be managed at different authorisation layers such as file system authorisation layer, SAS metadata authorisation layer and database authorisation layer.

35. Feedback will be transferred to the NSIs using secured channels.

36. The system components directly exposed to the public will be operated in a dataless way. Confidential data will only be handled inside a secured server environment.

J. Required changes

37. The remote access can be provided by the web solutions of the SAS BI platform. This way existing Stored Processes can be reused and directly executed from the WEB interface. Some changes are required to develop the signalling processes whereby the NSIs can signal the upload of the final version of their data. Considering the lengthy processes, asynchronous processing is preferred. The user initiates the processes but the results will be sent to them by secured e-mail.

38. The main advantage of such a remote system is that it can shorten the throughput time as the data providers can have feedback on their data right after the upload.

39. Data validation is simplified as the data providers use the same validation rules as Eurostat and those can be more complex than the rules of the data provider (e.g. comparison to European data available in Eurostat).

40. It will simplify the workflow and establish a closer link between the data provider and Eurostat.

K. General conclusion

41. Inside Eurostat, the GSAST platform is proven to work satisfactorily for the processing of several microdata collections. Challenges in handling the relatively heavy metadata structure have been tackled through the implementation of a specific GSAST Metadata Editor and this fosters increased independence of power users from the central support services.

42. Possible extension towards NSIs presents several challenges. While the proposed architecture ensures secure treatment of microdata, the adaptation of working methods presents a major challenge which is, however, fully in line with the joint strategy of the ESS.

References

Pongas and Wirtz (2011) Statistical data editing near the source using cloud computing concepts, MSIS 2011 Luxembourg.
SAS Institute, (2011) GSAST Feasibility study 2011. (Available at Eurostat on request.)