

Integrated Information Systems for Multilateral Statistical Processes: The Case of Purchasing Power Parities

Gregory Farmakis¹, Paulus Konijn², Maria Glossioti³

¹AGILIS S.A., e-mail: Gregory.Farmakis@agilis-sa.gr

²EUROSTAT, e-mail: Paulus.Konijn@ec.europa.eu

³AGILIS S.A., e-mail: Maria.Glossioti@agilis-sa.gr

Abstract

In the field of statistical information systems, research emphasis has typically been placed on concepts related to the integrity and representation of statistical data or computational problems, usually assuming however a simplified staged or linear model of the data flow, from data collection and validation to publication. In the case of multilateral statistical processes, involving several countries, this approach unavoidably leads to the collection and processing of distinct data sets through several resource consuming iterations. For complex multilateral statistical activities however, such as the estimation of Purchasing Power Parities (PPP), which present the additional challenge of continuous and coordinated collaboration of numerous national statistical institutes, the typical staged architecture, based on the mere exchange of data sets between the distinct data sources (i.e. countries) and a central data hub (i.e. Eurostat), would be prohibitive. The architecture and design of the new integrated information system for Purchasing Power Parities, makes extensive use of web accessible applications and autonomous XML conversant building blocks to facilitate a tight and continuous collaboration of experts from the participating countries, OECD and Eurostat during the highly interactive phases of survey design and preparation as well as of data validation. The design also required the effective synergy of several concepts such as history, concurrent versioning and multilinguality of data and metadata. The system offers a series of specialised collaboration mechanisms for humans and flexible data exchange for machines, transforming a tedious repetitive exchange of data sets into a continuous real-time collaborative process.

Keywords: Statistical Information Systems, Statistical Survey Systems, Statistical Data Exchange

1. Introduction

The architecture and implementation of typical statistical information systems are usually based on the underlying assumption of a somewhat oversimplified linear or staged model of the flow of statistical data. Data is supposed to flow in batches through the consecutive stages of data collection, data validation, processing, storage, aggregation and publication, which are assumed to be clearly delineated. At each stage data is usually extracted, processed using a specialised software module and

reloaded to a database. Thus research emphasis and technological state of the art have been primarily concentrated on issues mainly related to the modelling, representation and integrity of statistical data and metadata, such as metadata-based architectures, statistical data warehouse design patterns, confidentiality treatment etc. or computational problems such as outlier detection, while the potential of concepts such as real time collaboration, concurrent interactive statistical processing etc. has only relatively recently emerged as a strategic priority.

The same staged paradigm is usually applied in the case of multilateral statistical activities. In the case of multi-country data collection, following this linear concept, raw or aggregate data are usually collected and processed in batch mode in the form of distinct data sets that are collected, validated and processed. When iterations are needed in this process, usually during the validation phase, entire data sets circulate back and forth in a time and resource consuming fashion.

Data is collected by a network of sources, usually national statistical authorities in member states, and transmitted to a central data collection hub, usually Eurostat, in the form of files or even spreadsheet documents. Then data is compiled, validated in batch, and if problems such as outliers and missing or inconsistent values are detected, they are reported to the relevant source, which is expected to send an updated data file. On the other hand, the usage of validation methods and software tools at these sources, obviously, cannot eliminate the need for cross-validation of the entire compiled data set.

Thus, while this staged approach can be tolerated in the case of local, centralised Statistical Survey Management Systems, it proves costly and inefficient in the case of multilateral statistical activities.

Apart from the need for coordinated data validation, complex statistical activities, such as the estimation of Purchasing Power Parities (PPP), a truly multilateral exercise involving continuously evolving coordinated surveys, also present additional challenges: the need for effective, timely, almost continuous and coordinated collaboration of numerous national statistical institutes throughout the entire survey life-cycle, from the early survey design and preparation phases, to iterative aggregation – computation – validation cycles until the required level of data quality is achieved.

This paper presents the architectural principles and the technological strategy we implemented while designing and creating the new integrated information system for the PPP statistical activity, as well as the experience from the first period of its actual operation.

2. The Purchasing Power Parities Multi-lateral Statistical Activity

The Eurostat-OECD PPP Program is an ongoing activity since the early 1980s. Its purpose is to calculate PPPs for the member states of the EU, the member states of the OECD and third countries and to use the PPPs for comparisons of the GDPs of the countries involved. Eurostat is coordinating the collection of the necessary data within

each participating country, receives data from the countries and compiles and disseminates PPPs for them.

The calculation of PPPs relies on price and nominal expenditure data for a detailed breakdown of GDP, from the expenditure side, into specific goods and services. Final expenditure on GDP is broken down into a large number of detailed categories for which expenditure data can be obtained in all countries. The categories are further broken down into individual goods and services for which price data can be obtained. The calculation of PPPs proceeds from the bottom-up, starting with the prices of individual goods.

The participating countries are organised in groups and conduct regular surveys in order to collect data on the prices of consumer goods and services, prices of capital goods (equipment goods, construction), rents and salaries.

2.1 Collaborative Survey Design

In the case of PPP, needs for extensive and almost continuous collaboration arise early, in the survey design phase, where participating countries need to iteratively reach a common list of very specific items (i.e. goods or services) to be priced. There are ten distinct price surveys for each cycle of the process, six for consumer goods and services, two for capital goods as well as surveys for rents, and salaries. For each price survey a product list must be compiled that comprises an equi-representative selection of comparable products for each basic heading that is to be surveyed. Thus, a sample of products that can be priced over a number of distinct countries has to be defined, while these products must be comparable across all participating countries pricing them as well as be representative of each country's expenditure on the respective basic heading.

The process is based on the concept of an item list, including the exact products and services to be priced, as well as numerous detailed specifications for each one, so as to ensure comparability. In the case of consumer and capital goods, as well as services, participating countries have to conduct pre-surveys based on which the selection of items to be priced has to be finalized. During these pre-surveys countries ascertain the availability and representativity of products proposed for the product list as well as verify whether the products have been sufficiently specified to ensure comparability across countries.

During this process, proposals for deletions or additions of products, amendment of the specification of their characteristics, or inclusion of new products with their specifications will be discussed, amended, modified and accepted. The modifications continue when group lists are merged to ensure overlapping between groups. This may require the combination of products with similar characteristics while group leaders are able to propose new or modified product specifications to another group or modify their own group product list according to another group's list.

With 37 countries organized in four groups, coordinated by group leaders and four to five hundred products per survey, this process has to be effectively managed and

coordinated, following a common timetable, in order to ensure comparability between the country groups and avoid imbalances.

2.2. Collaborative Data Validation

Coordination and collaboration needs are also extensive during the data collection and the highly iterative validation phases, aimed to identify and eliminate non sampling errors and outliers from the survey price data.

National authorities in consultation with the group leader perform outlier checks on the prices of each product (intra-country validation), while the group leaders make outlier checks on reported prices comparing them with the corresponding prices of other countries (inter-country validation).

The validation procedure is inherently iterative: identified outliers are corrected and new average prices and PPPs are computed in order to identify more outliers. During the validation phase price observations of each group member are separately checked and edited using questions and answers between group leaders and group members. In addition, the average survey prices of each group member are checked against the average prices of other group members. Corrections are agreed and new validation-correction rounds are initiated until the quality of the data is satisfactory.

3. The new Integrated PPP System

Obviously, the typical statistical information system architecture, based on the mere exchange of data sets between the distinct data sources (i.e. countries) and a central data hub (i.e. Eurostat) would be inefficient. Thus, when designing and implementing the new integrated information system for Purchasing Power Parities, we decided to make extensive use of web accessible applications to facilitate a tight and continuous collaboration of experts from the participating countries and Eurostat during the highly interactive and iterative phases of survey preparation and data validation.

3.1. Collaboration Mechanisms

The novelty of the system however, extends well beyond the concept of allowing different users to concurrently work on the same data (which by itself presented several serious design challenges) by incorporating various collaboration mechanisms.

For instance, during the survey preparation phases, users can at any time post comments, questions and answers and a country can make suggestions, or adopt and adapt the suggestions of another country, concerning the specifications of a specific item, or even simultaneously compare the suggestions of several countries.

Similarly, during the validation phases, a country group leader may initiate a discussion concerning problematic data while countries have access to other countries' average prices for comparisons. This way, a slow procedure involving the repetitive exchange of data sets can be transformed into a continuous "real-time" collaborative process.

3.2. Architecture and Data Flows

Apart from these collaboration mechanisms, the architecture of the system is designed so as to enable a seamless process flow from survey preparation to publication of results. This presented a series of design challenges including a multidimensional data warehouse design with data versioning capabilities, maintenance of the history of item lists as well as of statistical observations, multi-linguality, elaborate access control, translation support mechanisms, as well as the use of XML and web services for data exchange.

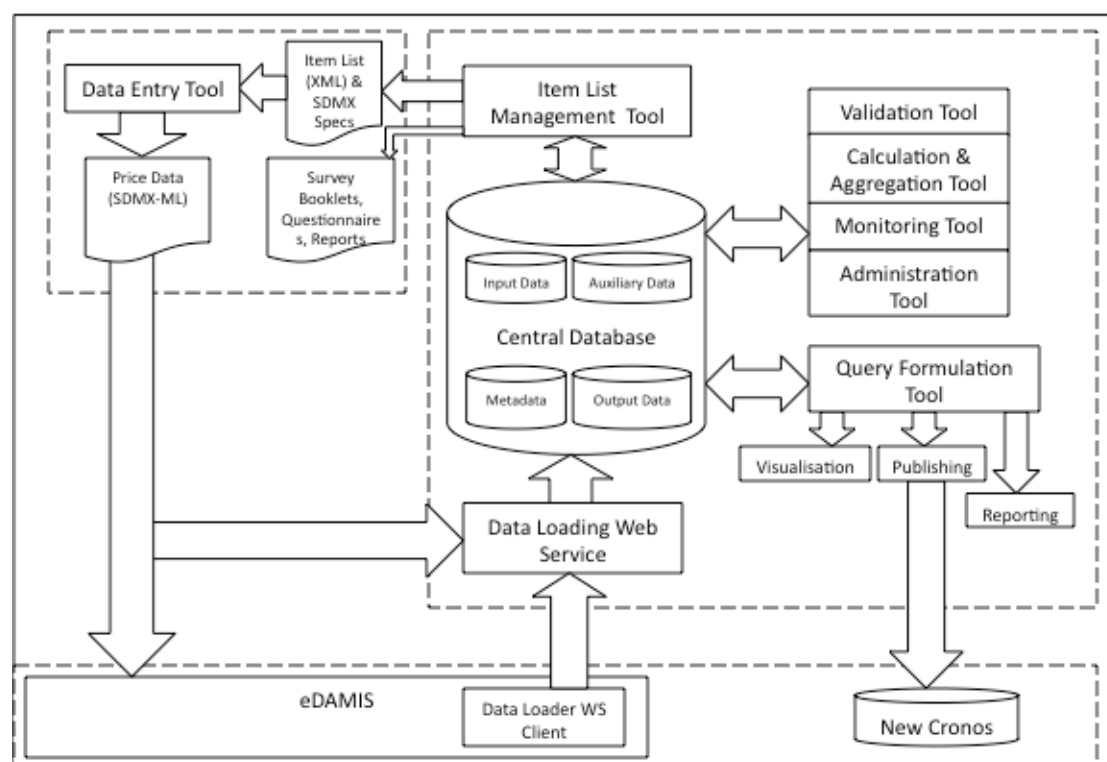


Fig.1 – Architecture and data flows of the integrated PPP system

The architecture of the integrated PPP information system (fig. 1) includes:

- the **Item List Management Tool**, i.e. a web application supporting the iterative and collaborative survey preparation process, including a metadata base with survey and item metadata;
- a **Data Collection Tool**, for the collection and initial intra-country validation of price data;
- a **Loader Interface**, i.e. a web service capable of receiving initial validated price data sets, validate them again and store them in a central data repository,
- the **Data Repository**, i.e. a multidimensional metadata based data warehouse of price data;
- the **Validation Tool**, i.e. a web application supporting the iterative and collaborative intra- and inter-country validation phases;
- the **Calculation Tool**, in charge of performing the calculation and aggregation procedures;

- the **Visualisation Tool**, i.e. a web application allowing multidimensional OLAP operations and publication of the data; as well as
- the **Monitoring Tool** supporting the coordination, time planning and monitoring of the entire process.

The Item List Management Tool, through which the survey preparation is accomplished, can also support the translation of the Item List, even including a thesaurus of item related concepts and suggestion mechanisms to facilitate the translation process. The tool can also deliver the translated final list of items and their characteristics to any data collection application, in a machine understandable XML format, thus reducing the burden for the preparation of data collection for the participating countries.

Furthermore, a specialized data entry tool has been developed, which is capable of automatically retrieving the applicable item list for the specific survey. This tool can, compile data from different price collectors and can also perform initial intra-country validation, prepare the initial data set and submit it to a data collection web service using the SDMX XML standard. Already, the mere use of XML and web services, by allowing syntax validation, ensures a first level of data quality, while digital signature and encryption mechanisms can be incorporated.

These data sets are validated again upon reception so that outliers are detected and “flagged”, and data is loaded in a multidimensional data warehouse, capable of maintaining multiple versions of the data. The web based Validation Tool allows then the collaborative validation and correction of data, this time for both intra- and inter-country comparisons, while the history of corrections is maintained. Acting on the same data warehouse the Calculation Tool can estimate and store versions of the PPP indicators, while the Visualisation Tool allows interactive multidimensional OLAP operations and publication of results.

4. Conclusions

The experience from the operation of the system proves that the extensive use of web-based collaborative mechanisms, not only improves timeliness and cost effectiveness of multilateral statistical activities, but furthermore, can also be a powerful catalyst of change towards the streamlining and re-engineering of these activities.

In the case of PPP, intra-country validation was delegated to group leaders who detected outliers, reported them back to the countries and waited for the updated datasets. Similarly, inter-country validation was done at the country group level only, by group leaders, before being completed at the European wide level, while these phases were conceptually and time-wise isolated.

With the new PPP data collection and validation tools, comparative information at the European-wide level is available as early as the datasets are collected. Intra-country validation and inter-country validation at the country group level and at the European-wide level can now run in parallel; countries can perform their intra- and inter-country validation tasks without delegating these tasks to group leaders so that the role of the

group leaders is upgraded from mere data quality control to quality assurance and consulting.

Similar conclusions hold for the survey design phase (i.e. item list management), where the isolation between distinct phases is now dramatically less rigid.

Thus, the impact of simple but efficient and specifically designed collaboration software mechanisms extends well beyond the mere quantitative improvements in terms of time and cost into the realm of qualitative process re-design.

References

- EUROSTAT-OECD (2006) *Methodological Manual on Purchasing Power Parities: European Communities – OECD*
- Sundgren, B. (1995) *Guidelines for the modelling of statistical data and metadata*, Conference of European Statisticians, United Nations, Geneva 1995
- Sundgren, B. (1999) *An information systems architecture for national and international statistical organizations*, Invited Report, Meeting on the Management of Statistical Information Technology UN ECE Geneva, Switzerland, 15-17 February, 1999
- Farmakis G. et al (2001) *Architecture and Design of a flexible Integrated Information System for Official Statistics Surveys, based on Structural Survey Metadata*: NTTS 2001