Statistics Portugal Survey Management System Architecture

Luísa Pereira[1], Luís Ferreira[2]
[1]Services Director, Software Engineering Unit, Statistics Portugal, e-mail: luisa.pereira@ine.pt
[2]Analyst Developer, Software Engineering Unit, Statistics Portugal, e-mail: luis.ferreira@ine.pt

Abstract

The Survey Management System Architecture is the result of a great effort by Statistics Portugal to concieve, integrate and modernize its own production system.
The main goals were:

- Implement an integrated and consistent approach to survey management systems;
- Integrate, harmonize and reuse processes;
- Increase coherence and comparability by using the same updated population;
- Reduce statistical burden on data providers.

Due to the complexity of the task, a divide and conquer technique was used, and the system was approached by identifying a set of smaller subsystems. Each was addressed individually, keeping in mind the integrated nature of the overall target system.
Although a unique integrated information system was designed, responsibility for each of its subsystems was distributed by several departments.

Keywords: Integration, Sample design, Data collection

1. Introduction

The Survey Management System began operation in 2008 and is the result of a very important and hard work undertaken by Statistics Portugal to concieve, integrate, modernize and put in practice its own production system.
The first survey integrated into the Survey Management System was Intrastat. At the moment, there are about fifty surveys integrated, corresponding to the majority of business surveys and in 2013 all business surveys will be integrated in this system.
In early 2006, the conception of this system was started by defining its main goals.
The whole system was designed in-house by a cross-domain workgroup, with a single leader appointed by the board of directors, with well defined deadlines and common goals.
Before describing the system itself, it is important to discuss the triggers (reasons) that made us realise we actually needed an integrated system and were confident we could do it. A presentation of the system and its subsystems follows. Finally, culminating with a

global evaluation on the work done, some considerations on the lessons learned and a brief glance at what is planned for the future.

2. The triggers - Background and context

Survey management in Statistics Portugal has been, for many years, defined and documented in a process called “Statistical Production Process Manual” (SPPM). All participants involved in statistical production must act according to roles and responsibilities described in this manual. However, different domains work in slightly different ways and the implementation of the processes defined in the manual takes different forms depending on the “spirit”, “tradition” and needs. So, Statistics Portugal had a traditional stovepipe approach with multiple vertical and compartmentalized requirements.

In terms of software development to support this work model, the same route was followed, and distinct systems were put in place for each statistical domain. The software was developed according to the mindset of the specific people who were going to use it.

This work model, when dealing with surveys on business statistical units, starts by creating a population which is fetched and frozen from the Business Register (FUE) to be used for one year.

FUE gets updated with fresher information via administrative and other sources, while the population doesn’t change.

From the population, a sampling frame and sample are derived for each survey. Samples are handed as a file to the production team responsible for that survey, who inserts them in the specific user software.

Once the sample is on the specific software, business statistical units are updated locally over time.

To further illustrate, suppose the same business gets selected for two surveys (in the same time period). During data collection, data about that business may be updated in each survey. This results in redundancy and possibly incoherence.

In the end of the data collection process, a file is created with the business data gathered. This file is then sent to the FUE’s responsible, so it may be used to update business statistical units.

Updating FUE is a difficult task when dealing with redundant, possibly incoherent and late data. To make matters worse, context data (paradata) is missing. It’s very hard and time-consuming to choose the right data to merge into FUE.

Production users tend to disfavour FUE data and use their own updated version, furthermore, they tend to neglect the work of sending fresh data to FUE as they don’t recognize any advantage in doing so, leading to even more outdated data in the main FUE repository.

Next year’s population will be of lesser quality than it would be if the above mentioned problems could be avoided.

Since production experts were used to work in “their way”, it was hard to get everyone to see beyond the horizon of their specific needs and goals and to think together to achieve a higher, more efficient goal.

In face of this “compartmentalized” environment, the information architecture used in Statistics Portugal was a pool of small disparate information systems that didn’t “talk” to each other. Another shortcoming was that the “information architecture” was not geared towards information but instead, towards the roles people have in the processes.
Statistics Portugal wasn’t ready to meet the requirements society demanded:

- Businesses (data providers) complained that information they sent, more than once, was being discarded;
- Businesses (data providers) complained they were being asked too often to answer to an excessive number of surveys;
- It was impossible to have an accurate measure of the statistical burden on businesses, and attention to reduce statistical burden has been gaining relevance lately;
- The same task done by different people had different requirements and outcomes.

Change was unavoidable!
Statistics Portugal had to come up with new solutions that would fully answer the new demands presented.
The required change was too big and complex to be carried out all at once.
Statistics Portugal was aware of the difficulties inherent to big changes. Many people are afraid of change, afraid of being rendered useless, afraid of losing control, afraid of losing power within the organization, and finally, others simply don’t see the need for it.
The urge to solve the redundancy, inconsistency and lack of freshness of business statistical units data was unbearable.
The answer came through a “window”!
Integration was done in a “non-intrusive” way, by giving users the perception that user applications were completely decoupled. To accomplish that, a “window” was created which allowed any user application to “peek” at FUE and interact (access and update) with it, ensuring the accountability of all changes.
The FUE “window”, invoked from user applications, provides data about business identification, localization, characterization and control (time and source of last update of each attribute).
Some of the attributes provided by the window, mainly related to identification and localization, are updated directly on FUE’s repository.
The remaining attributes are updated via the submission of update proposals. After examination, a FUE expert accepts or rejects the proposal and justifies the decision to the proponent.
Additional control data (paradata) was put in place to allow auditing and attest the transparency of the whole process.
Every time a user application needs some business data (identification, localization...), it will get it from FUE repository instead of using their data models.
New updating rules for business statistical units were set, allowing for all users to access and update, resulting in a higher quality FUE.

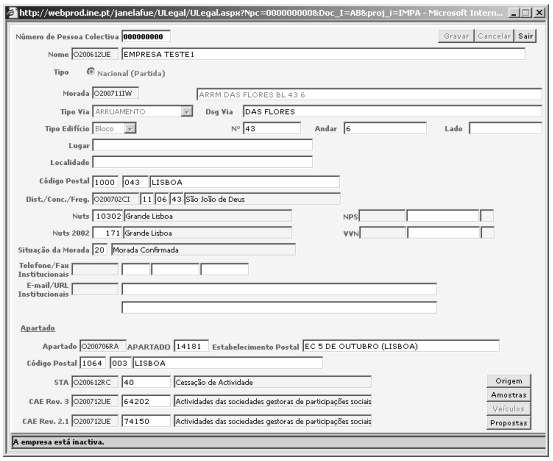


Figure 1: The FUE "Window"

The FUE "window":

- solved the problems of redundancy, incoherence and lack of freshness of business statistical units data, with undisputable gains in the quality of the population;
- proved that integration and interaction between systems was possible;
- helped break the resistance to change and allowed to build high hopes for the other changes that needed to happen.

Shortly after the FUE "window" was deployed, an organizational restructuring took place resulting in the creation of a central Data Collection Department to handle all data collection and the creation of three domain production departments (Economics, Social and Demographics, and National Accounts). Methods and Information System were merged into a single department.

This reorganization reinforced the need for streamlining the data collection process and led to a new SPPM in 2010. A mapping can be established between the new SPPM and the Generic Statistical Business Process Model (GSBPM) (Saraiva, P. 2012).

Summarizing:

- FUE "window" triggered the collapse of the resistance to change and raised expectations on the creation of an integrated informational system in Statistics Portugal;
- Working on being able to "measure the statistical burden" triggered the insight that only through integration is it possible to measure and control;
- The "centralization of roles" triggered the standardization and reengineering of data collection process.

Everything was in place to start the design of an information system architecture from an integrated and information-oriented point of view.

The architecture was outlined, presented, and met with the approval of the board of Statistics Portugal, which formed a cross-domain workgroup to design, develop and supervise its implementation.

The active participation of everyone was fundamental, but the commitment from the board of directors was paramount in the successful materialization of such a system.

3. Survey Management System

The Survey Management System (SIGINQ) is an integrated environment designed to attain high-quality statistics production. The main features are: a dynamic population; a dynamic and centralized repository of sampling frames and samples that finally allow to effectively measure the statistical burden and improving the process of sample design;

generic and modular software that takes advantage and promotes the standardization and reuse of processes.
To build such a complex system, a divide and conquer technique was used, and the system was approached by identifying a set of smaller subsystems. Each was addressed individually, keeping in mind the integrated nature of the system intended.
The system wasn't built from the ground up. Instead, according to best-practices of software engineering, some legacy systems were improved and reused.
Although a unique integrated information system was designed, responsibility for each of its subsystems is distributed by several departments: Production, Methods and Data Collection.

Figure 2: SIGINQ subsystems

3.1 Business Register

Business Register (FUE) is a Statistical Unit Registers System for businesses.
FUE repository is refreshed via administrative sources and data from surveys.
FUE was one of the reused systems. It was modernized to accommodate new statistical units related to business – such as Local Unit, Road Vehicle and Periodical.
The concept of "window" had been a recent major improvement to FUE.

3.2 Population and Sample Management

Population and Sample Management System (SIGUA) has an integrated database where the annual population for each kind of statistical unit, related to business, can be found.
Each survey has its own sampling frame and sample, but all the surveys having the same kind of statistical unit, have the same annual population.
Since all sampling frames and samples reside in the same database, burden measurement becomes feasible and this measurement greatly influences survey design.
In SIGUA focus was put on sample design and tracing down the statistical unit life cycle.
SIGUA has its own update process, independent from FUE.
The concept of "window" is reused. In data collection management software a "window" is available which allows interaction with the SIGUA system:

- A data collection/production expert updates the statistical unit data by creating a change proposal;
- The proposal life cycle breaks down to: analysis; decision; respond to proponent with the result of the proposal. In this cycle, production, FUE and methodology experts are involved;
- If the proposal is accepted, sample, sampling frame and annual population are updated. FUE experts may decide to update FUE repository.

Each statistical unit type is characterized by a set of variables, which can be of two kinds:

- Up-to-date variable – must always be up-to-date and contains the "current" value;
- Time-based variable – reporting to a specific period in time and contains the value at a given time. These are also known to be characterization variables.

When accessing up-to-date variables, the value is retrieved from FUE.
When accessing time-based variables, the value is retrieved from sampling frame (SIGUA).
Identification and localization are up-to-date variables (like name, address…).
Variables used in stratified sampling are time-based variables (like Statistical classification of economic activities-NACE, Nomenclature of territorial units for statistics-NUTS, …).
One variable present in FUE and in SIGUA may be collected in a questionnaire, if so, there may be different values for the same variable.
The system keeps all those variables' values and the update process allows users to update exactly the value they intend to.
Since a variable can have different values depending on where its value is retrieved from, it is very important to understand and internalize this concept, which is represented in figure 3.

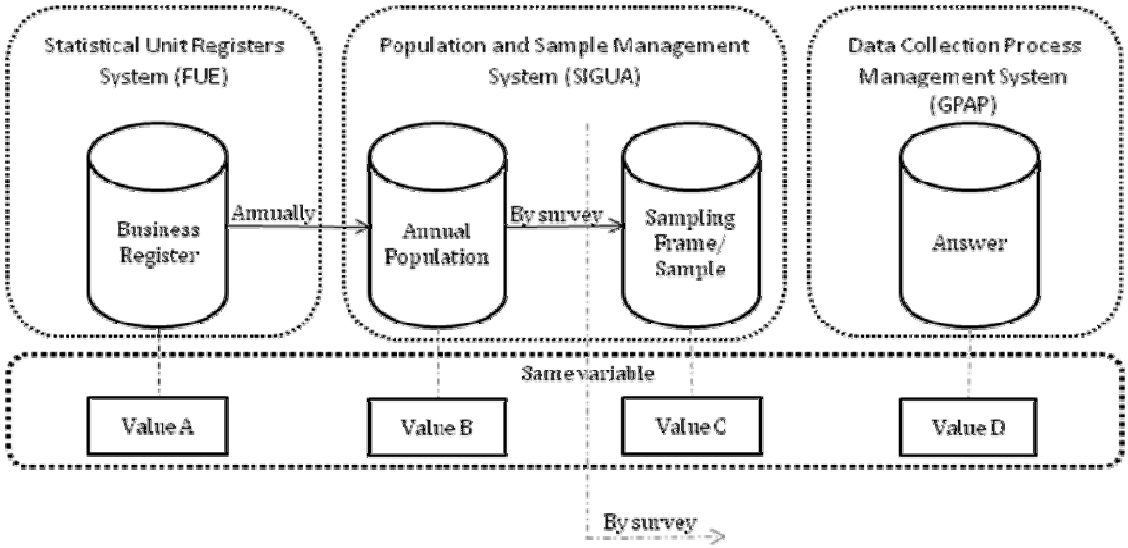


Figure 3: Statistical unit values

This was a ground-breaking improvement with positive implications for survey methodology, of leaving "compartmentalized" approach behind and embracing an "integrated systems" approach to statistics production.

3.3 Respondents Management

Respondents Management System (GRESP) intends to optimize respondent cooperation and encourage their participation, following the recent trends on Customer Relationship Management (CRM).
There are two kinds of respondents:

- WebInq members – respondents who use WebInq (Statistics Portugal internet site where respondents can submit their answers);
- Non WebInq members – respondents that respond to Statistics Portugal by paper.

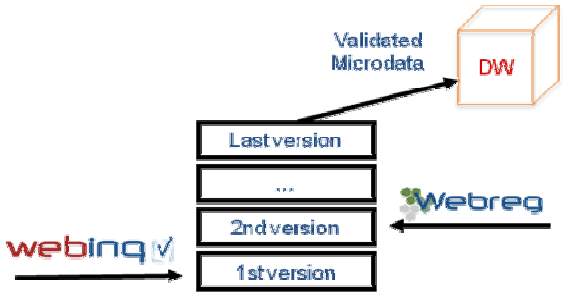
All information about respondents is maintained in this system and shared across all subsystems. It's always up-to-date, due to the updating rules defined for respondent data.
Due to the impact WebInq had in reducing the workload of post-collection team, a brief overview of WebInq is in order.
The Internet service WebInq is in production since 2005 and has currently 55 surveys, 113 416 registered users and 161 810 registered businesses. The number of valid answers submitted in 2012 was 835 770.
The difference in the number of between registered users and businesses is due to the fact that each registered user may submit answers on behalf of one or more businesses.

3.4 Data Collection Process Management

Data Collection Process Management (GPAP) is the pivotal system. It receives samples from Population and Sample Management System (SIGUA), adds up-to-date statistical unit variables from Business Register (FUE) and combines them with respondent's data (GRESP) in order to launch the survey.
Data collection process has several kinds of operations to be conducted. Some may apply to several surveys whilst others may be specific to just one. This requirement was addressed by creating standard and harmonized common processes without forgetting an area where to include specific processes.
Each time a new survey enters the "Survey Management System", an assessment is made to determine if there are any specific needs. If so, further analysis takes place to see if that specificity can be regarded as a new generic functionality, extending the system's range of available features. This flexibility is achieved through survey parameterization and sensible default values.
For illustration purposes, let's consider a menu that appears on screen in the software that supports GPAP. The survey menu options available respond to parameterization.
The deliverable from the data collection process is a set of validated microdata. This validated microdata is then sent to the data warehouse where it is subject to further processing and deeper analysis.

3.5 Data Capture and Editing

Data Capture and Editing is supported by two components: Webinq and Webreg.
Webinq, questionnaires on the web, is the component where respondents submit their answers online.
Webreg is the component which allows data collection teams to post-collect paper-based answers (usually sent by traditional mail or fax) or edit answers.
Data Capture and Editing System is mainly interconnected with GPAP.
Data capture and editing is a time-consuming activity and critical to data quality. Acknowledging its relevance, a new concept was created: "Data Version Stack".



Data Version Stack has four main rules:

- Single response for each tuple (survey, data reference period, statistical unit);
- Data editing performs a new response version;
- All versions of the responses are saved;
- Last version goes to data warehouse.

Figure 4: Data Version Stack

Data Version Stack allows the survey manager to track changes and ensure data quality.

3.6 Call Centre

Call Centre was the last implemented system. It was designed and built after all the other pieces were in place and took advantage of all the lessons learned and features available. Otherwise, it would have been unthinkable to implement such infrastructure.
To help survey manager to follow-up with the respondent, Call Centre provides two channels: telephone and mail.
Interactions between Statistics Portugal and respondents are context aware and tracked. In business domain, the most useful features are telephone reminders and support to respondents.

4. Integrate a new survey in the Survey Management System

The first thing to do, when integrating a new survey in the system, is to assure that it meets the necessary requirements.
Once the requirements are met, it's time to start feeding the various subsystems:

- Metadata experts have to insert it in the Statistics Portugal Metadata System;
- Methodology experts have to register data sampling method in Population and Sample Management System;
- IT analysts have to introduce Data Collection Processes parameters.

The main issue is to identify wherein the survey needs, which ones are common and which are specific. Specific needs have to be specified and coded.
To help prepare specifications, a set of normalized documents and templates was created where experts can specify parameters for common processes, as well as describe specific processes. This time saving starting point is titled “Guide to integrate surveys in SIGINQ” and usually referred to as “the Guide”.
With the help of “the Guide” and given that all tasks are now simple and standardized, all stakeholders have an easier time integrating a survey into SIGINQ.

5. Conclusions

Building an integrated information system and all the supporting software for a complex system with high organizational impact as this one is a rewarding and exciting task.
Looking back it was surprising to find that some small and apparently self-contained achievements unintentionally triggered big system-wide changes and helped push the system forward.
Building large scale systems in small steps is better than a “big bang” approach.
Choosing in-house development brought undisputable gains in terms of the project’s cost and management. The know-how of an in-house cross-domain team, the full support from the board of directors, and the commitment from the leaders of the members of the workgroup were crucial during the whole process.
On the downside, this approach made it hard to accurately define the scope of the system and to negotiate changes while sticking to the planned schedule.
Productivity of software development team increased significantly. This productivity is measured by the number of surveys deployed and the time it takes to integrate a new survey.
Productivity also increased in production and data collection as design, specification and test tasks were largely reduced. Integration of a regular survey on the system is simply a matter of defining parameters, questionnaire design and validation rules.
Human resource management in Statistics Portugal also benefited in three ways:

- Reallocate human resources from one survey to another at a fractional cost;
- In-house training and support time have been greatly reduced;
- Less risk of single point of failure in human resources.

This system is a reality, its benefits are tangible and noteworthy, nonetheless there is still a long journey ahead in extending, optimizing and handling more and more of the statistic production workflow.
Having an integrated information infrastructure gives us a strong belief that Statistics Portugal will be able to respond properly to the future demands of Society.

References

Saraiva, P. (2012) Integrated Data Collection System on business survey in Statistics Portugal, European Conference on Quality in Official Statistics – Q2012