Centralize Data Collection: Does it worth?

Paulo Saraiva dos Santos1, Almiro Moreira2

1Department Director, Data Collection, Statistics Portugal, e-mail: paulo.saraiva@ine.pt

2Unit Director, Data Collection, Statistics Portugal, e-mail: [almiro.moreira@ine.pt](mailto:almiro.moreira@ine.pt)

**Abstract**

Statistics Portugal (INE) started in 2004 an integrated and process driven approach aiming at reengineering its production, improving the efficiency and flexibility.

Consequently, a central data collection department was created, regional directorates were extinct, and domain production departments have been merged into three units: economics, social and national accounts. Methods and information system were merged into one department.

Survey’s data collection is a core function of Statistics Portugal, consuming around 40% of its annual budget and 30% of its human resources. A Data Collection department assures mainly the production phases of collection, processing and analysis of collected microdata, covering all business and social surveys.

The authors consider the centralization of the data collection process was facilitated by a broad internal reorganization at that time.

The benefits have been substantial, especially on innovation and quality improvement. Nevertheless, it remains many opportunities for further developments.

This paper describes a historical view of this journey and lists the results achieved. In addition, it offers some general guidance to evaluate if centralize data collection processes is a good choice for the national production of official statistics.

Key words: Central Data Collection, statistical production management.

1. **Introduction**

The production of official statistics is a complex chain of processes, and faces increasing challenging realities. While society wants diverse and faster results, data providers resist collaborating, and respondents are also demanding for advanced modes to provide their required data. Other recurrent issue is the decreasing funding of official statistics, but we also have new opportunities through the technological advances. [1] [2]

Although there is no ideal way to organize a national statistical office (NSO), there are evidences of high level of similarity in the broad models adopted by different countries. Thus, improving the exchange of experiences between peers is a clever way to cope with the challenges.

Statistical offices with a high level of dependence on surveys have their data collection activities between the most resource consumer functions. Consequently, having data collection centralized in a single unit is a frequent option to be taken into account because of its potential of benefits; one is the improvement of the relationship with the respondents. But there are also disadvantages to consider, especially transforming a highly decentralized data collection structure.

The authors aims at this paper sharing ten years of experience of Statistics Portugal in implementing a centralized approach to data collection, improving its efficiency and flexibility and contribute to the discussion on the centralization of the data collection processes within a NSO.

In most countries, there is more than one producer of official statistics at the national level. In order to focus at the NSO level, this other dimension of discussion is not considered here.

A statistical office uses to focus on two ways of structuring its organization: by subject matter and by function. While a subject matter unit stands for looking into statistical themes and domains, such as Business and Social statistics, or National Accounts, a function unit is mainly based on knowledge, skill, work process or business function. It is always difficult to make a clear cut choice between both ways. Thus, the structure of most statistical offices is a mix of the two.

Considering the GSBPM Model [3], this paper stands Data Collection as most of the sub-processes from the *collect* and *process* phases. Data collection is said to be centralized when all or most of the sub-processes of these phases are assured by a single organizational unit, no matter if its resources are geographically centralized or not.

It is very common that data collection activities are conducted by individual survey units, using a “silo mode” approach. Nevertheless, there are many examples where the data collection activities have been progressively centralized, like Australian Bureau of Statistics, Statistics Canada, and Statistics Portugal.

There is no single recipe to follow, but there are some essential ingredients to be considered. Like any other organization, a Statistical Office needs to continuously understand its particular context, evaluate its performance (real and perceived) and organize its resources and processes accordingly.

As the context is in constant change, centralize or decentralize functions like Data Collection have the tendency to act like the movement of a *pendulum*, varying from different degrees of centralization. This is perfectly acceptable, and even expected, when a NSO correctly evaluates its performance of the way it complies its mission (processes, resources management and organization).

1. **Data Collection centralization at Statistics Portugal**

**2.1 Background and context**

Established in 1935, Statistics Portugal is the Portuguese central authority for the production and dissemination of official statistics, which ensures the supervision and the technical and scientific coordination of National Statistical System. Being mostly a survey-oriented statistical office, data collection is a core function of Statistics Portugal, consuming around 40% of its annual budget and 30% of its human resources.

Like many other NSO, prior to the year 2004 Statistics Portugal produced statistics through a non-integrated organizational architecture, based on numerous parallel processes, domain by domain, place by place. Data collection activities were conducted by individual survey departments and regional directorates. A survey organizational unit would design their survey, but the development and operations were usually not harmonized. Centralized activity was mainly restricted to some software development and to a weak coordination of the household interviewer workforce.

**2.2 Organizational evaluation in 2003**

After being appointed in 2003, a new administration board decided to make an extensive organizational evaluation in order to better understand the existing reality and identify opportunities to improve Statistics Portugal’s performance. By that time, the production and dissemination were assured by the headquarters (Lisbon) and a number of regional offices located, respectively, in the North, Centre, South-East and South of the country. Together, the five Regional Directorates accounted for a significant percentage of the total staff of Statistics Portugal, and their main activities covered the following areas:

* To assure local data collection, considering that proximity to respondents were more effective;
* To coordinate at least one full subject matter project with national scope, considering that each region has its socio-economic specificity and it had advantages to better produce the corresponding statistics;
* To develop regional statistics, considering that these were better handled out than the headquarters because of its proximity to the users and sources.

Firstly acting as dissemination and data collection center for the region (NUTS II), the regional offices gradually assumed active role in statistical production and regional studies, and its organization and resources have increased fast.

Concerning the former recruitment policy, it resulted into imbalanced capabilities between regions and Lisbon. Most of hiring occurred in the regions during the period from 1990. Because of that, the average age of the employees outside Lisbon is well below the total average. And since that hires had a better educational level, there is of a much higher proportion of staff with higher levels in the regions than in Lisbon. This fact has limited substantially the alternatives to reorganize the office.

The final diagnosis was very insightful:

* Unbalanced level of geographical decentralization, with a strong dispersion of the activities;
* Rigid microstructure with a high level of subdivisions with small dimension, inhibiting the sharing of resources;
* Data collection organized by survey and scattered by thematic areas and the regional directorates, having an insufficient level coordination, resulting in a burden on respondents and duplication of the required information;
* Dispersion of responsibilities of the support departments;
* Absence of effective management systems;
* Limited ability to change and adapt to the evolving information needs of society.

**2.3 Restructuring process (2004)**

With the evidence of such inefficient and not flexible way of producing statistics, Statistics Portugal launched an intensive restructuring process. A project to re-engineer the production architecture was undertaken based on an integrated and process driven approach aiming at improving its efficiency and flexibility.

The reorganization was based on the premise that improved performance should be done without substantially increasing the resources available, resulting in a significant increase in efficiency.

Consequently, a central data collection department was created, regional directorates were extinguished, and domain production departments have been merged into three units: economics, social and national accounts. Methods and information system were merged into one department.

Coordination of projects began to be made by the departments of matter in Lisbon, which came to focus the activities of design, analysis and integration of information.

It was a remarkable challenge, considering the new distribution of resources, roles and responsibilities. The transition was carried out in a way that the statistical operations were not substantially affected, in spite of some resistances and other constraints.

**2.4 The Data Collection Department**

Today, a Data Collection department assures mainly the operation of statistical production phases of collection, processing and analysis of collected microdata, covering all business and social surveys.

Data collection staff is spread all over the country (mainland and islands), especially in Lisbon, Oporto, Coimbra, Évora, and Faro, but under centralized system.

The Autonomous regions of Madeira and Azores have their own authorities for the production of regional specific statistics, while being the data collection centres for those areas for Statistics Portugal, under common technical requirements and infrastructure

The Data Collection Department has now local centers, where data collection by home interviews continued to be managed based on the geographical location of each local centre, albeit with centrally defined methods and procedures. The collection via other modes (postal, telephone and electronic) became distributed by local centres, not by geography but by thematic area, in order to leverage existing human resources in local centers and enable the specialization of functions and benefit from economies of scale.

Having around 200 employees and 350 interviewers, Data Collection is organized into three units:

* Self-completed surveys, assuring the data collection activities related with business surveys, which are mainly supported by self-completed questionnaires;
* Interview surveys, assuring the data collection activities related with social surveys, which are mainly supported by interview methods (CAPI and CATI);
* Data Collection Processes, assuring the promotion and modernization of tools and processes to support the collection and analysis in the context of geographical information systems and paradata development, as well as CAPI (face-to-face), CATI (telephone) and CAWI (web) processes. This unit also assures de centralized management of CATI Surveys.

**2.5 Data Collection management**

The management is decentralized but centrally controlled, fostering high levels of individual autonomy with a clear definition of responsibilities. The objectives are defined in “cascade”, from the Department, to the units, to the sections and, finally, to the workers.

There is one overall budget distributed through each management level.

The Activity Planning cycle is conducted as the following:

* Statistics Portugal Annual Activity Plan;
* Data Collection biannual activity plan;
* Evaluation of human resources needed;
* Project distribution between the teams;
* Some human resources are shared between two units;
* Preparation of Service Level Agreements (SLA) proposals to subject matter departments.

There is a matrix management of the human resources, which means that the resources can be allocated accordingly to the each organizational unit’s portfolio.

Concerning the field management, Statistics Portugal uses only freelance interviewers to social surveys at the households and price collection at the stores. Interviewers have annual contracts and their wage is variable, been calculated according to the work done. Interviewers can assure CAPI, CATI and price collection. Field work is supported by laptop and other technological infrastructure, assets owned by Statistics Portugal.

CATI interviewers work at home (HomeCATI), using a common and secure infrastructure with online technical support by a central team, which allows freelance interviewers work at home, integrated in a virtual contact centre and based on a voice over Internet protocol (VoIP) solution. Innovative solution with many advantages, but there are challenges, especially planning schedule, case management, strategies with available resources, interview supervision and monitoring.

Statistics Portugal has an Interview Management System, which supports all the processes related with social statistics and the price collection. This tool has several components: team management and the tools used by the interviewers to collect data, transfer them to Statistics Portugal, allowing them to work both in face-to-face and telephone interviews.

**2.6 Responsibilities of Data Collection and Subject Matter**

The division of work between Data Collection and Subject Matter was based on an internal reference model to describe the statistical business processes at Statistics Portugal, called “Statistical Production Process Manual” (SPPM) which is comparable to GSBPM.

Data Collection is responsible for the *Collect* phase and part of the *Process* phase (5.1 Integrate data; 5.2 Classify & code; 5.3 Review & validate).

Dealing with the relationship between Data Collection and Subject Matter was one major challenge at the very beginning.

There was a negative perception of the data collection function, seen as “a low profile work …”. Conversely, subject matter statisticians were very “data collection oriented”.

There was always a very high expectation. Often there were discussions like this: “… in my time, I proceed in this particular way. If you do not collect the data as the usual, you will have to prove that you are doing better. “

The solution was implementing Service Level Agreements (SLA) to manage expectations and to build trust. At the beginning, we try not calling it SLA. The term has a negative perception and it does not help the negotiation. It was used a step-by-step approach, from a simplified version and increasing gradually the complexity. Nowadays, SLA is a core solution widely accepted internally.

SLA are project specific, having two main components:

* Basic data collection level, which is fully taken as granted and is included in the methodology of the survey;
* Variable data collection level, which is negotiated during the preparation of the survey.

SLA essentially contains operational indicators, with a complete definition of its measures, goals and intervals of acceptance and to overcome. These indicators are continually measured and treated as paradata.

**2.7 Integrated Production System**

From the early stages of the reorganization it was clear to the Data Collection team that the benefits of the centralization of the function would only succeed if it was supported by an integrated production system.

With the support of the administration board, a multidisciplinary working group was established, composed by representatives from information systems, methodology, statistical domain specialists and data collection. The mandate was focused on designing an integrated architecture to support the production.

This effort resulted in an Integrated Survey Management System (SIGINQ), which covered firstly the business surveys, and later on the social surveys.

From 2008 until now, the number of features supported by SIGINQ grew up continuously. Nowadays, all of our surveys are supported by this system. [4]

**2.8 Electronic questionnaires and automated data collection**

The first result after the centralization was a significant improvement of the electronic data collection for business surveys. In July 2005 the Internet service WebInq was launched, offering an easy and secure alternative to businesses to provide the survey data through electronic self-completed questionnaires. Nowadays, more than 85% of the business questionnaires are collected electronically.

Improved by a Eurostat Grant, Statistics Portugal started an automated data collection project in 2012 with very successful and encouraging results.

This complementary data collection method allows data providers to submit in a single XML file, the variables related with multi-statistical units, multi-period and multi-surveys.

Using automated data collection, a data provider can collect data in their own information systems, generating a XML file and submitting it to Statistics Portugal pushing a single button, and avoiding manual filling in into a web form. Data providers save time and resources, reducing their reporting burden. On the other hand, Statistics Portugal benefits improved data quality and faster response time.

**2.9 Improved communication with data providers**

Having an integrated production system is a key element to improve the relationship with respondents. As long as the data collection processes occurs, paradata is automatically collected, offering an extensive knowledge of the data providers and respondents.

Nowadays, it is easy to answer some questions like: What are the surveys that a company or a respondent is involved? What is their response behaviour? Those questions where near impossible to answer with the former approach.

Harmonized skills are used to provide an improved capability for respondent communication. Standard reminder and follow-up strategies are in place, having an active capability to modify the implement approaches accordingly to the development of a certain data collection campaign.

Written questions, complaints and suggestions from respondents are handled centrally to ensure a consistent reaction based on standard templates.

Statistics Portugal efforts to improve the consistency of respondent communication are illustrated by one example. In 2013 Statistics Portugal carried out this ad hoc survey to data providers, which allowed obtaining a picture of how the obligation to reporting statistics is viewed by the respondents. With the results was possible to have further information about the Perceived Response Burden and the quality of the data reported, and obtained critical suggestions for the improvement of procedures in data collection.

**2.10 The way forward**

Statistics Portugal remains seriously engaged with the modernization of data collection and further developments are planned for the near future:

* Extend the coverage of WebInq to household statistics, providing CAWI surveys. In September 2014 Statistics Portugal will launch its first CAWI survey: the European Health Interview Survey;
* Extend the automated data collection to all business surveys, as an alternative to self-completed questionnaires;
* Develop SIGINQ new features, such as the improvement of the analysis of paradata, and extension of score analysis model to more surveys;
* Invest in systems to collect regular transactional data, such as supermarket scanner data;
* Evaluate the adaptations needed to extend the use of the SIGINQ to other statistical authorities;
* To use the data collection structure and culture as enabler to reduce the survey-dependence of the Portuguese official statistics production.

1. **Conclusions**

* Considering the experience of Statistics Portugal in centralizing data collection in the last decade, there are substantive results as well as many opportunities for future modernization and transformation;
* The economies of scale and the productivity gains delivered by the centralization of data collection have freed up resources to invest in more innovation and to develop new capabilities;
* When having a data collection team focused and motivated, centralize data collection functions is an important enabler for the innovation that is required for a National Statistical Office. This approach allows new corporate capabilities to be developed relatively quickly.

Consequently, for Statistics Portugal it worth the enormous effort to centralize the data collection functions. Nevertheless, we would like to repeat that there is no single recipe to follow, but there are some essential ingredients to be considered. Like any other organization, a Statistical Office needs to continuously understand its particular context, evaluate its performance (real and perceived) and organize its resources and processes accordingly.

1. **References**

[1] Bruengger, H. (2008). How Should a Modern National System of Official Statistics Look?

[2] Edmunds, R. (2005). Models of Statistical Systems. PARIS21, Document Series 6.

[3] The Joint UNECE / Eurostat / OECD Work Session on Statistical Metadata (METIS) (2013), Generic Statistical Business Process Model v5.0 (GSBPM),

[4] Saraiva dos Santos, P and Valente, C (2012). Integrated Data Collection System on business surveys in Statistics Portugal. European Conference on Quality in Official Statistics. Athens. Working paper and presentation.