When do you hear the warning bells?

Pedro Cunha (pedro.cunha@ine.pt), Statistics Portugal/Information Infrastructure

Sónia Quaresma (sonia.quaresma@ine.pt), Statistics Portugal/Information Infrastructure

Jorge Magalhães (jorge.magalhaes@ine.pt), Statistics Portugal/Information Infrastructure

Abstract

Quality information can make the difference between good and on time decision making and support, and poor or delayed results. In order to increase data coherence, a strategic data warehouse has been built by Statistics Portugal in the last decade. Several challenges were faced not only on methodological issues but also concerning data owner and stewardship, management and generally promoting the reuse of already collected information. All these steps improved our data quality and got the warning bells ringing whenever there is a sign that there may be a problem. But sometimes the alert did not arrive soon enough and we needed to put in place an early warning system. Due to the data warehouse exploration, a lot of expertise on business intelligence was acquired and so Statistics Portugal decided to take a step further and use the capabilities and in-house know-how to follow the data collection surveys since their first moment - the field work. Promoting data quality from the beginning of the collection process was our challenge for the last couple of years and we now propose to share the experience we accumulated during this process.

Keywords: Promoting the collection process quality; early warning system; data quality; business intelligence; data coherence.

1. Introduction

In the last decade the Portuguese NSI underwent several organizational changes, a large part of which affected the way the surveys were carried out. While some modifications were carried out having efficiency objectives in mind, to diminish the burden upon the respondents without compromising the overall quality of the survey was always a concern. At the same time that the survey budget and schedule had to be more closely controlled also the response rate had to be evaluated in almost real time.

Those issues were the basis for the establishment of new monitoring goals concerning our surveys.

At the same time and keeping in mind the NSI apprehension about compromising the collected data quality, several reports and template assessments were conceptually designed to ensure the checks and evaluations any particular survey should follow.

During the instantiation of the above mentioned models with any given survey we understood that to keep an eye on the collected data quality meant observing the data and sometimes even scrutinizing it.

In this paper we'll make a short overview of the Portuguese NSI organizational changes and explain our motivation in using BIS to achieve four main goals:

- evaluate survey's response rate.
- keep budget and schedule survey controlled.
- make the survey first results quickly available for the statisticians and methodologists.
- perform daily data quality assessments.

This paper is organized as follows. Sections 2 and 3 describe Statistics Portugal reality concerning business intelligence solutions and how its internal organization has created opportunities for enhancing the use of BIS. Section 4 focuses on the advantages of using BIS at the various stages of the statistical process. Conclusions are made in Section 5.

2. Statistics Portugal reality

For the last fifteen years Statistics Portugal has been using data warehouse to analyse and disseminate data and to also improve data quality. In this process, the human resources have gained extensive knowledge in the use of web intelligence tools, by building, sharing and disseminating their own ad-hoc queries.

The warehouse approach provides the means to store data once, but enabling its use for multiple purposes. A data warehouse treats information as a reusable asset. Its underlying data model is not specific to a particular reporting or analytic requirement. Instead of

focusing on a process-oriented design, the repository design is modelled based on data inter-relationships that are fundamental to the organization across processes [3].

Data warehousing became an important strategy to integrate heterogeneous information sources in organizations, and to enable their analysis and quality. This happens because data from external sources is cleaned, transformed, aggregated and integrated to provide more accurate and useful data to the user. In this process, data quality is assured by:

- analyzing out of range values: whenever values that are not allowed in a given dimensions are detected, the record is examined by checking its quality;
- analyzing missing values: if a dimension has no value and should have, this must be reported and further examined;
- identifying outliers: if, for example, it is expected that a measure has at most 1000 units and if the record is 10 000, the situation must be evaluated.

When analyzing data, end users can detect multiple quality problems such as:

- large variation of values between periods;
- lack of consistency in data. For example, in the agricultural census, it is asked how many acres of hybrid maize the holding has and how much of that maze is watered. If watered hybrid maize acres in a certain region is greater than total hybrid maize, this means that this information is not correct;
- abnormal correlation between two or more variables of the survey. For example, if the agricultural census data tell us that, in a certain region, there are 10.000 cows, it is not possible that there are only 10 stables;
- analysis of data from a survey crossed with data from other surveys. For example, if the agricultural census data mentions that a certain region produces 50.000 litres of milk, that information must be compatible with the result of the milk collection survey for the same year.

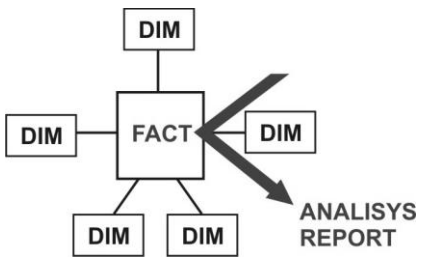

Figure 1 – Data warehouse schema

For these reasons, data warehouses in National Statistical Institutes (NSI) are considered a very powerful tool when dealing with large amounts of data as they allow for the improvement of data quality. However, sometimes, problems are discovered too late, leading to a time consuming and expensive process to solve them. For that reason, often, we hear the warning bells too late.

On the other hand, we have On Line Transactional Processing (OLTP) databases that support survey management. It is well known that these databases are very powerful responding to data manipulation as inserting, updating and deleting, but are very ineffective when we need to analyse and deal with a large amount of data. Another constraint in the use of OLTP is their complexity. Users must have a great expertise to manipulate them and it is not easy to understand all of that intricacy.

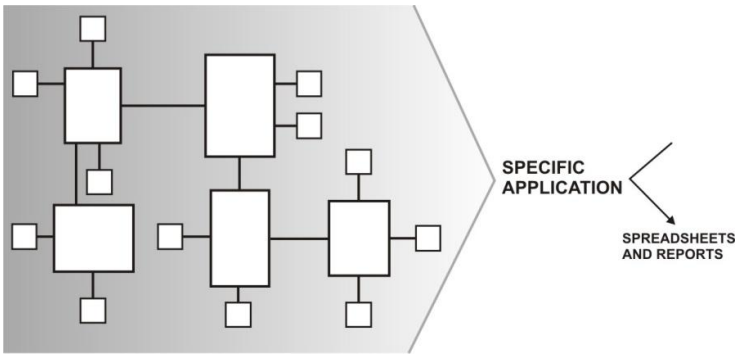

Figure 2 – Relational Database Schema

Statistics Portugal whenever a survey is released a specific application is created to support that operation. Usually, for validation purposes, internal users have access to a set of twenty spreadsheets that are created according to their specifications. Whenever users require other frames or changes to the existing ones, an intervention of the IT development team is required. That process is time consuming and on early IT response may not be as efficient as needed.

However, these databases have, much sooner in the process, the information needed to analyse the data and promote its qua1ity. They have the fundamental microdata for that analysis and all the management information that data warehouses do not provide.

Therefore, why not consider an intermediate level that inherits the best qualities of both systems? Why not think over a business intelligence solution (BIS) that has microdata and fields necessaries to improve data quality using already familiar tools. These tools provide simplicity when working with data but are powerful enough to deal with large amounts of data and have the ability to quickly aggregate data.

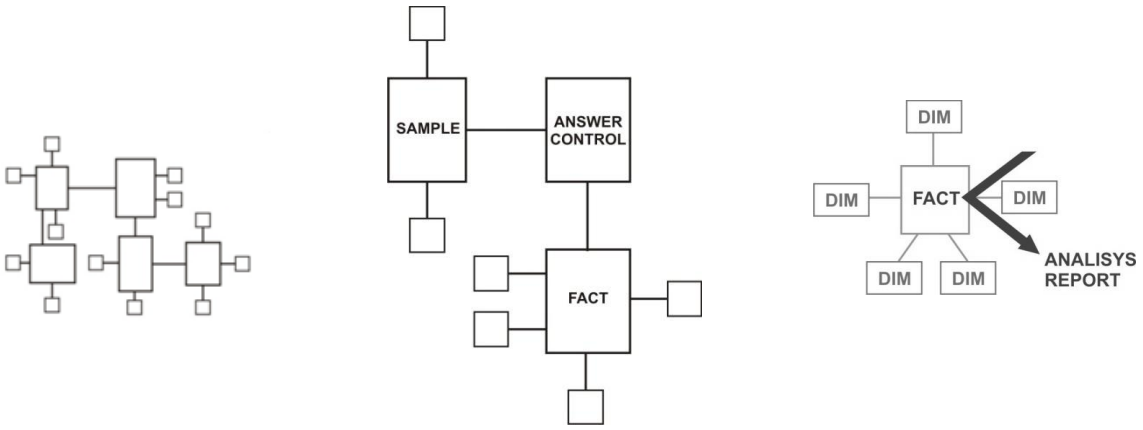

Figure 3 – BIS in the middle

3. Changing the structure of an organization can boost the use of BIS

In 2005, Statistics Portugal made a major change in its organizational structure modifying significantly the way different departments use statistical data. These changes were implemented for organizational purposes but also to respond to the new vision on the production methods of European Union statistics.

Before 2005, each department had all the responsibility for its own surveys with managing power over the entire production process – from field data collection to data dissemination; they oversaw all the stages of the process taking the necessary steps to improve the quality of data at every point.

In that changing year, an important restructuring of the internal organization was made. A centralized department called Department of Data Collection (DDC) was created with the responsibility of gathering the data of all Statistics Portugal surveys.

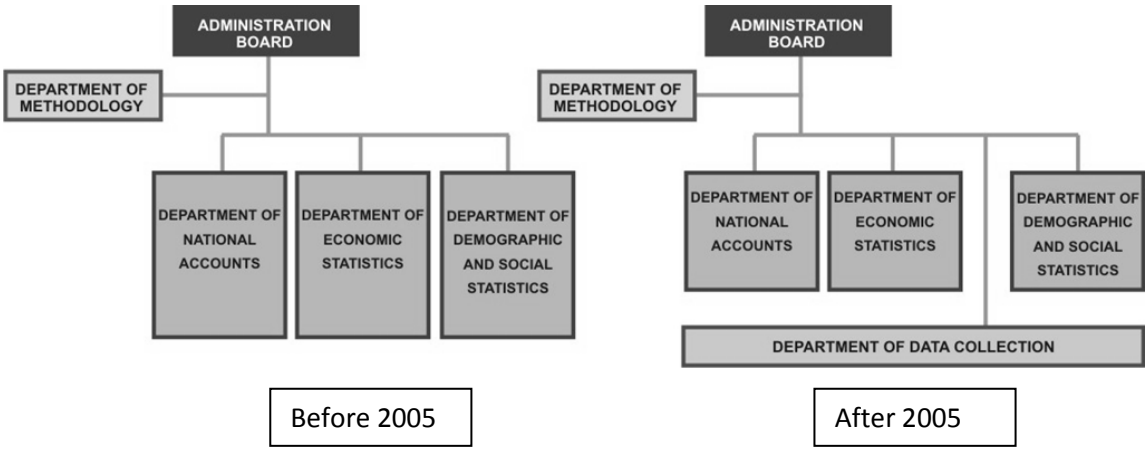

Figure 4 – Statistics Portugal simplified structure before and after 2005

DDC collects data using different techniques depending on the survey type: Direct collection – using his interviewers' team across the country; postal and web collection for enterprises surveys; and more recently Computer Assisted Telephone Interview (CATI).

After the restructuring, the three statistical units responsible for the surveys - Department of demographical and social statistics; Department of economic statistics; and Department of national accounts - kept the control over the management and the dissemination of information but lost access to collected data until the validation by DDC, after imputation, estimation or stratification of underlying data if necessary, is completed.

Before this change, access to data was much focused on narrow statistical areas. For example, DDC accessed data through dedicated computer software while statistical units accessed data warehouse through business intelligence tool. However, the three statistical units expressed the wish to access the data earlier in the process. They wanted to assess data quality, field work evolution, response rate and preliminary results. Statistics Portugal decided that it was a very good opportunity to extend the use of its business intelligence tools, well known by their technicians, to data collection.

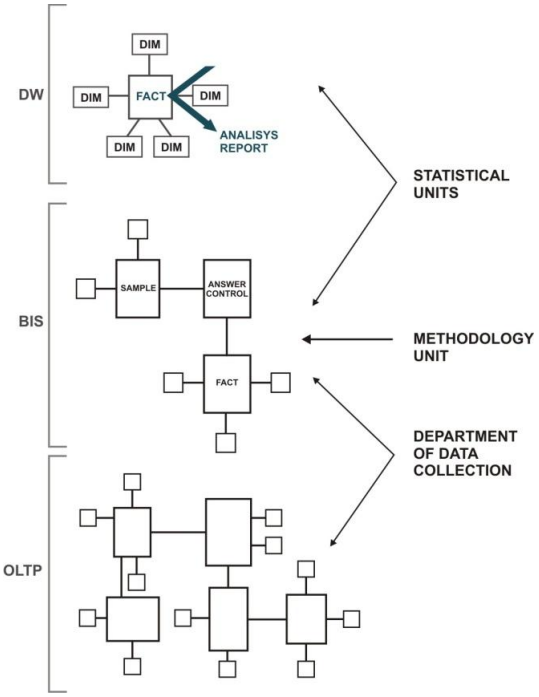

Figure 5 – Departments and access to data

4. Why BIS?

A number of trends common to both industry and government are coming into play. They are helping to make BI pervasive and to fulfil the promises made by BI vendors of rendering BI accessible to the masses. Trends include moves by industry and government to implement, where appropriate:

- operational and embedded BI;
- real-time BI;
- mobile, geo-aware, and social BI.

While BI had previously been applied only in the analysis of historical data collected in data warehouses and data marts, operational and embedded BI bring BI functions to the front line, to everyday, line-of-business applications.

Operational BI provides monitoring and alerting capabilities, and it may provide the ability to react in real-time, with low analytical latency to current (and not just historical) conditions. BI with low analytical latency, where results are returned near-instantaneously whether involving historical or operational data or both, is termed real-time BI.

These applications have been promoted to real-time execution, and they take advantage of modern data systems architectures to run against fresh, up-to-the-moment data [4], as freshly collected survey data for instance.

Our BIS has to be accessible to all users, easy to use, always up to date, must have information about the survey sample and answer control and must have management information.

We built our BIS having in mind:

Accessibility: Any technician that has access to our intranet and have one browser can access any BIS he has been given permission to use. The platform to access BIS and data warehouse is one and the same.

We also provide specific areas in our wiki corner with premade reports in excel. These reports have on-line information about survey data but also management information. This enables managers at various levels to have easy access to all survey using familiar tools.

Usability: Using business intelligence tools already known by NSI workers eases the extensive use of BIS. As BIS is very simple and intuitive to use, individuals very easily build their own queries and share them with others.

Using survey sample: By including survey samples in BIS, users have the possibility to analyse the answerer's information at any time like the name, contact, economic activity classification, and other meaningful information.

Answer control: In our BIS, we include information about answers from each respondent for the entire period they have to answer the survey. For example, if it is a monthly survey, it is possible to analyse if a respondent answered in one month but did not in another one or if he did not answer at all. We can study the evolution which is very useful not only to issue early warning but also to discover patterns, for example in the labour force survey.

Management information: It is possible to extract information that helps several aspects of survey management. Aspects like who answered and when; what is the answer status; who included the respondent in sample.

4.1 Using BIS to manage surveys

An essential key to create BIS is the possibility to have survey management.

Reports and spreadsheets with information that helps the data collection are easily made. It is possible to create queries that show reports for:

- how many interviews ended with success, how many were unsuccessful, when they were made it, how they were made;
- how many records each interviewer (directly in the field or by phone) made;
- how many refuses each interviewer got, thus detecting training necessities;
- how much time was spent in each interview or in all interviews;
- the survey response rate;
- the satisfaction questionnaire.

All these measures can be crossed with various dimensions like time, geography, interviewer, employer, respondent, supervisor, economic activity classification, etc.

Another role for BIS is his capability for managing payments to external workers. Before, payment to interviewers was a big problem. Interviewers were paid per interview, so, when interviews were completed and inserted, payments were carried out. However, in many cases, those interviews were incorrect and DDC had to reopen that record. Interviewers were asked to cross-examine that interview. However, they did not have much incentive to do that because they had already been paid. Construction of BIS helped this management in the sense that DDC has more flexibility in controlling field work.

4.2 Using BIS to improve data quality

BIS allows users to analyse variables in a thorough way. They can make some of the validations they do in data warehouse but much sooner using well known and easy to use tools. As it was said earlier, it is possible to assess large variations of values, the consistency of data and the correlation between variables and between surveys. By doing that we are making the “warning bells” ring whenever we have a sign that there may be a problem.

But what about other “bells”? Is it possible to make other validations to data with this new system?

As BIS has on-line micro-data, it is possible to make the same validations we did before but now at the respondent level or aggregating using some of the dimensions we will have available in the data warehouse.

It is also possible to make other quality analyses like value absence, abnormal values in variables or poor data quality.

mês	Chicken	Turkeys	Ducks	Quail	Ostriches	Geese
02	13.783.093	580.535	294.407	**14.658.035**	248.197	693.580
03	12.364.295	234.628	216.786	**12.815.709**	295.308	693.341
VAR% MA	-11,47	**-147,43**	-35,81	**-14,38**	15,95	-0,03

Num	01	02	03
	Cow's milk (l)	Cow's milk (l)	Cow's milk (l)
499263884	100.209	82.457	85.601
499291950	46.692.749	44.430.250	48.316.845
499372160		1	1
499372179	13.608.004	13.058.841	14.309.771
499836663	1.540.186	1.523.567	1.634.192
499852561	669	661	749

Figure 6 – Quality analysis examples

We can get response rates for geographical or other stratifications used by the survey, and to react in real time if some stratum is having poor answers or not getting them at all.

We also built our BIS with some derived variables that have validation purposes: variables calculated in order to serve the interests of data collection department, statistical units and methodology unit.

Taking labour force survey as an example:

Based on the interview's start and end time, a new classification divided into three classes: "Long Interview", "Short Interview" or "Not Obtained" is calculated. This variable is used for payments to interviewers (long interviews are better paid).

Other variables are built with alerts based on the path of the answers given, that is, even though answers are correct, their chaining could not be correct.

Statistical units use this derived variables: if, for example, someone goes from an employment situation to an unemployment situation, and that person has under age children, an alert is issued to the interviewer to ask if that person has backup support. On

the other hand, if he has children over sixteen that already work, other questions could be asked about their current situation.

Another example of derived variables used by these departments is, for example, for how many years the individual lives in Portugal or if he is getting paid by any professional activity, with the new derived variables being built from variables asked in the survey.

The methodology unit also uses variables that are calculated in BIS. Until BIS was constructed, this unit received in each quarter a text file to calculate all variables needed for stratification and calibration. When BIS was built, all strata of various variables (like payment, economic activity, geography, number of children, etc.) became automatically calculated.

REGION		MONTHS 02			MONTHS 03		
		SAMPLE	RESP.	RR%	SAMPLE	RESP.	RR%
11	Norte	25	25	100,00%	24	24	100,00%
16	Centro	29	29	100,00%	28	25	89,29%
17	Lisboa	18	18	100,00%	17	16	94,12%
18	Alentejo	29	29	100,00%	29	28	96,55%
20	Região Autónoma dos Açores	36	35	97,22%	36	35	97,22%
30	Região Autónoma da Madeira	7	7	100,00%	7	7	100,00%

Figure 7 – Response rate by region

5. Conclusion

Statistics Portugal faces a permanent challenge to produce quality statistical data in the most efficient way, considering both budgetary and time constraints, and without increasing the burden upon the respondents. These issues were the basis of the establishment of new monitoring goals concerning our surveys.

In this paper, we explain how the Portuguese NSI organizational changes turned out to be an opportunity to enhance the use of business intelligence solutions and how these solutions are effective in achieving the above-mentioned goals. BIS combines the best of two worlds: some of the capabilities of data warehouses (DW) to deal with large amounts of data, and the ability of On Line Transactional Processing (OLTP) to manage data manipulation.

The use of BIS at Statistics Portugal was well accepted by the workers involved and ended up by improving the efficiency of data collection through a better management of the field work and the creation of additional variables. The development of BIS has also boosted the use of DW since statistical records became cleaner and easier to upload.

At present, we are trying to extend our BIS in order to account for the dependency between individual data and also working on the refinement of our lodging databases to better support CATI's activities. To conclude, the experience of using BIS at Statistics Portugal has been extremely positive and we expect to keep on enhancing the efficiency and effectiveness of our statistics by using BIS.

References

[1] Quaresma S.: A brave new world. International Conference on Quality in Official Statistics. Mainz, Germany (2004).

[2] Vassiliadis P.: Data warehouse Modeling and Quality Issues. Ph. D. Thesis. 2000.

[3] Radermacher W., Baigorri A., Delcambre D., Kloek W., H. Linden: Terminology relating to the implementation of the vision on the production method of EU statistics. Presented in European conference on quality in statistics. 2010.

[4] Grimes S.: Government Business Intelligence (Space Time). http://www.spacetimeresearch.com/government-business-intelligence.html accessed on 15-04-2012.

[5] Sorce A.: The impact of data warehousing on the management of statistical offices. Instituto Nazionale di Statistica (INSTAT).