

The New IT Environment for the Italian Consumer Price Survey

Riccardo Giannini, Federico Polidoro,
Anna Maria Sgamba, Marco Silipo, Fabio Spagnuolo, Antonino Virgillito
Istat, Italy, e-mail: rigianni@istat.it , polidoro@istat.it, sgamba@istat.it,
silipo@istat.it, spagnuol@istat.it, virgilli@istat.it

Abstract

The current organization of Consumer Price survey in Italy is based on a non-integrated approach where data is collected either on paper or on portable PCs and handled by a software written in COBOL. Data transmission to Istat is performed by loading sequential files on a web site. After the Municipal Offices of Statistics (MOS) have loaded micro data, Istat recovers them on its own archives and after a detailed data check, edits micro data. The current process is not integrated because Istat archives and MOS archives are physically separated entities: therefore intervention carried out on Istat archives has to be repeated on the MOS archives.

It is evident that the use, though partial, of paper questionnaires and the archives duplication are sources of non sample errors due in particular to measurement errors. This makes it more and more difficult to accomplish in an efficient way the increasing quality requirements in terms of accuracy coupled with timeliness and coverage.

A project currently being carried out at Istat has the aim of realizing a in-depth reengineering of the whole Consumer Price (CP) survey IT environment, in order to reduce the possible sources of non sampling errors and improve the possibility to measure data quality through different set of indicators. The new IT architecture is based on a centralized relational database that stores all the survey data. The new data collection application, running on mobile PCs, allows data collectors to load data directly into the database right after they are collected. The new control and correction application allows to perform checks and editing on micro data directly on the database, with no confusing data redundancy. Moreover it makes available different set of indicators to monitor data quality. In this paper we describe in detail the new IT architecture that will support the survey and the improvements in data quality expected from its introduction.

Keywords: Consumer Prices data quality management, Computer Assisted Data Collection, Distributed Architectures

1. Introduction

Consumer Price (CP) survey in Italy is being carried out starting from 1927. Survey evolution has gone together with the evolution of the technological environment for the data production process. In 2008 CP survey at territorial level is carried out mostly by 84 Municipal Offices of Statistics (MOS) that collect monthly data for about 400.000 elementary quotes that are processed in order to produce the CP indices both

national and harmonised. For the time being, elementary quotes are collected either by paper questionnaires or portable PC and afterwards they are registered by MOS on stand alone archives. Then after a preliminary data check, elementary quotes are sent to Istat adopting a predefined record track and Istat recovers data on archives on which it carries out the procedures for data treatment and elaboration: micro editing treatments carried out by Istat have to be repeated by each MOS on its own archive with all the consequent problems of alignment between Istat and MOS archives.

Istat is implementing a very important project for the reengineering of the IT environment supporting CP survey, allowing a significant improvement of the data quality management. According to the project goals, firstly, the data collection by MOS is carried out entirely through portable PCs furnished of a dedicated software (P1), reducing the possible measurement errors due to paper questionnaires and to data entry activities. Secondly, the architecture of the system is based on a centralized Oracle RDBMS, hosted at, and managed by Istat, allowing to eliminate redundancy of data among Istat and MOS, thus avoiding the possibility of mistakes due to replication and enabling for a more agile survey process.

The central database is accessed through the two software components: the system is structured into the data collection application (P1) and the control and correction application (P2). Both subsystems are developed in Java. Design and development of whole system has been carried out by the IT personnel of Istat. In general the centralized database architecture and the real-time updates allows for timely monitoring of data quality and data collectors activity, so any critical issue that can occur during data collection can be quickly detected and addressed.

In this paper, after a brief overview of the context of the CP survey, we describe the new IT architecture and the expected improvement of data quality it can provide.

2. The Italian Consumer Price Survey

2.1 Overview of production process

Consumer Price Indices (CPI) are produced in Italy by grouping together elementary data collected in the survey coordinated by Istat. Istat regularly calculates and disseminates monthly figures on CPI, either national (CPI for the whole nation and CPI for the households of blue and white collar) or harmonised ones (Harmonised Index of Consumer Prices, HICP). The survey on Consumer Prices is carried out both at central and territorial level. Istat collects data concerning prices of products that do not show any variability along national territory (cigarettes for example), that are technically too complex to be collected at territorial level for issues of quality adjustment (mobile phones for example) or of products whose consumption is not strictly linked to the territorial areas where data collection on the field is carried out (in particular tourist services). For the remaining products (in terms of weights 80 per cent of the basket in 2008) data collection is carried out by Municipal Offices of Statistics (MOS) that are officially in charge of data collection in the field.

Concerning data collection carried out by MOS and making reference to month t , the production process is schematically articulated in the following steps (into brackets the subject of the step):

1. days 1 – 21, month t : data collection in the field (MOS)

2. days 1 – 24, month t : data entry or data downloading from portable PCs (MOS)
3. days 1 – 24, month t (for MOS whose indices are used for the flash estimates data check), days 1 – 30, month t (for the remaining MOS): data check (MOS)
4. day 25, month t : data upload by MOS whose indices are used for the flash estimates (MOS)
5. day 26 – end of month t : data check concerning MOS whose indices are used for flash estimate (Istat)
6. end of month t : flash estimate (MOS, Istat)
7. day 1, month $t+1$: data upload by remaining MOS (MOS)
8. days 2 – 9, month $t+1$: data check concerning remaining MOS (Istat)
9. Mid month $t+1$: dissemination of definitive data (Istat).

All the above process is supported on the IT level by a set of software components developed at Istat using the COBOL language. One component is used for data collection in the field, while another handles all the archives, providing functions related to data correction and (until 2005) indices estimate. Copies of these pieces of software are available at Istat and at all the MOS.

2.2 Data quality issues

The steps listed above allow to better focus the attention on the main critical aspects in terms of quality of the present CP survey data production process. They can be resumed as it follows:

- The high possibility of measurement errors due either to the adoption of paper questionnaires or to data entry activities.
- The risk of non sampling errors for the very short time that is available for Istat (and for some MOS) to check and revise micro data.
- The general inefficiencies due to the archives duplication. A big effort is always required in order to avoid mistakes due to replication and to maintain consistency among copies after editing.
- The risk for punctuality and timeliness for the bottlenecks that characterize the data production process.

Besides these main critical aspects it has to be added that the present IT system, that manages the entire production process, does not allow to measure data quality, with the exception of some basic indicators.

The new IT architecture for Italian CP survey aims to deal with the above aspects on one hand improving CP data quality in terms of three aspects of Eurostat vector of quality (first of all accuracy and then punctuality and timeliness) and on the other hand allowing the measurement of data quality through different set of indicators, as we detail in the following section.

3. The New IT Architecture

3.1 Architecture Overview

Though the current software has been successfully used to support the survey since second half of 90s, the lack of a real IT architecture, in addition to the data quality issues mentioned above, poses general problems concerning the overall management

of the IT support to the survey that are more and more pressing. One key issue is related to maintenance. Since the IT structure of Istat has finally embraced relational databases and languages such as Java and PHP as the main choice for development, the support of COBOL in terms of new developments and training of new resources has been abandoned. Then, there is a need for converting all software which was written in COBOL to guarantee that all the software supporting critical activities can be maintained at any time.

Starting from 2005 an Oracle relational database is used for storing all survey data. Since the beginning of 2007 CP indices elaboration is also performed by means of Oracle stored procedures. Though this partly has solved the obvious security issues deriving from storing archives on local PC filesystems, the other problems can be tackled only by creating from scratch a new IT architecture that centers on the relational database, and is capable to manage the whole survey process in a seamless way, allowing to definitively dismiss any piece of software written in COBOL, with an expected big impact on the overall quality of data, as we detail in Section 4.

The development of such a new IT architecture requires an in-depth reengineering of the survey process itself, and new development of all the software. At the same time, this sort of radical intervention poses a big challenge in a context where a continuous monthly data production has to be guaranteed, so the various development phases and the switch off to the new architecture have to be carefully planned. Moreover, new software tools require training for all the actors involved. As a part of the migration project, training courses are being carried out by Istat, involving all MOS personnel.

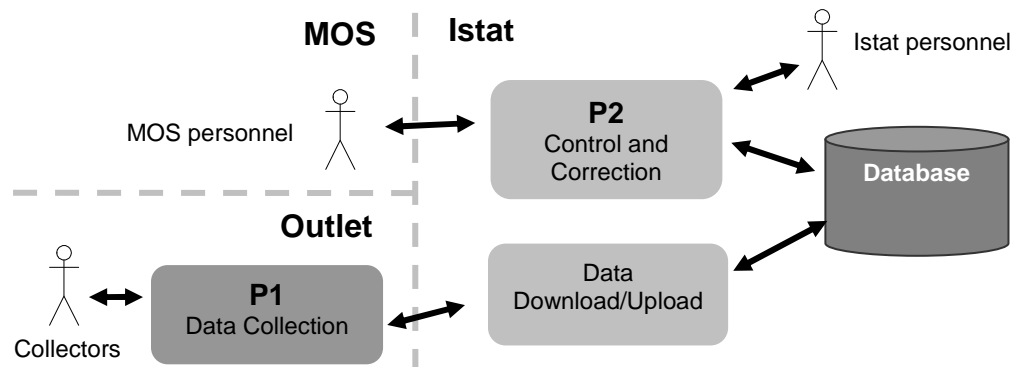


Figure 1: The new IT architecture

The new architecture is entirely based on Java technology and is depicted in Figure 1. It is composed of the following three macro-components:

- The central database: Oracle database hosted at Istat. Stores all the survey data and contains the software procedures for elaboration of indices.
- The data collection application (P1): Java desktop application running on mobile PCs provided to collectors. It features a touch-screen user interface with forms for price collection. P1 requires an Internet connection only for data transmission: once data is loaded inside the application the whole collection activity can be carried out offline.
- The control and correction application (P2): J2EE web application accessible by both MOS and Istat. It is mainly based on forms through which collected data can be inspected and corrected. It also includes functions for organizing and monitoring the daily activity of collectors.

A further components is required to integrate P1 with the central database: it is realized as a part of the P2 web application, accessible through the HTTP protocol and devoted to the download/upload functions of data.

We detail the general interaction of all the components while describing the new survey process in the following section.

3.2 The New Survey Process

We consider the operations performed by all the actors involved in the process along a month t , highlighting the usage of the various components in the architecture and their interactions:

1. *In the last days of the month $t-1$.* Personnel in all the MOS use P2 to assign collection tasks to all the collectors. This is written in the database.
2. *During the first 21 days of month t .* Collectors daily download data related to their assignments in their mobile PCs, using P1. Then, they visit the outlets they are assigned to for performing data collection, still through P1. Collected data is then uploaded to the server and it is immediately available for MOS to display and correct. MOS can change the assignments at any time during this phase. For example if a collector is temporary unavailable for illness her assignment can be moved to another collector that only has to repeat the download operation.
3. *Between 1 and 25* (for MOS whose data are used for flash estimate), *end of month t* (for remaining MOS) Istat and MOS monitor data production process and data quality, MOS edits through P2 the collected micro data.
4. *From 26* (for MOS whose data are used for flash estimate), *from the beginning of the month $t+1$* (for remaining MOS) *to the release of definitive data* (mid of month $t+1$), Istat and MOS monitor data quality, Istat edits through P2 the collected micro data .

The fact that data is available at MOS almost in real time means that in case a mistake is detected by an indicator it can be promptly corrected returning immediately in the field. This is not possible in the current survey organization, in particular when elementary data are collected through paper questionnaires. In general we expect a significant optimization just from the fact that all the actors will not be charged for data transmission tasks anymore, thus avoiding a time-consuming and error-prone activity that will be completely managed by the system.

For the sake of data consistency, during phases 3 and 4 locks are active on data so that concurrent modifications made by MOS and Istat are not possible.

3.3 The Data Collection Application

P1 is a desktop application, entirely implemented using Java 1.6. It has been developed according to the hardware equipment each collector will be supplied with, which will be a portable computer falling in the UltraMobile PC category. UltraMobile PCs have been chosen since they suit best the purposes of the application, providing the best trade off between weight and screen size, where lighter and smaller devices mount screens whose size is too small for comfortable operation with the

application, and proper laptops are a too heavy weight to bear for the data collectors. The computer used in the development phase is equipped with a 7" 1024 x 600 LCD touch screen display, it runs Windows XP Tablet PC Edition operating system and features a Intel A110 processor, 1Gb DDR2 RAM at 667 MHz and a UMTS card, providing ubiquitous mobile connection to the Internet.

All P1 functions have been designed and tested for use with a touch screen display with a resolution of 800x600 pixel. A large effort has been put on the optimization of processor resource usage and on the speed of the GUI frames, designing them as singleton Java Swing classes and adopting the Apache Derby as local database, since it has a light footprint, it is embedded and open source. Interaction with the DB is implemented through the widely used Hibernate framework.

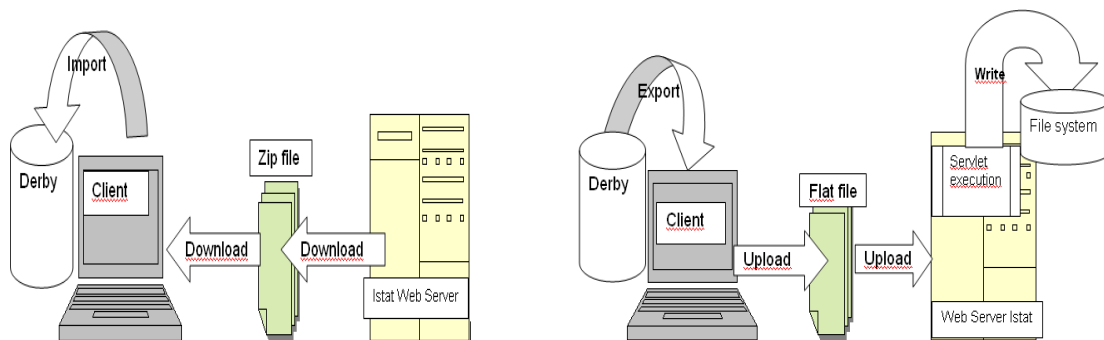


Figure 2: Client-Server communication (data download and upload)

3.3.1 Client – Server Communication

Client-server communication exploits the HTTPS protocol. Quotes are downloaded from the centralized server controlled by Istat. This activity is performed in three steps (Figure 2 – right side).

1. A unique zip file containing the list of quotes for the month (represented in flat text files) is downloaded from the server.
2. The zip file is unshrunk and deleted.
3. Data from the flat files is stored in the local Derby database

Once data is stored locally in the DB, quotes collection is carried out. At the end of her daily job, the collector sends back to the centralized server all the data he collected so far, in the form of flat files. The upload phase is done by means of a Java Servlet available on the Istat web server which is invoked for each single file. It simply copies the file on the server's filesystem, in a folder which is specific for each collector (Figure 2 – left side).

3.3.2 User Interface Design

The P1 application features a user interface that is able to meet the peculiar usability requirements of this scenario.

P1 user interface has been derived from the widely used and mature previous COBOL version. Several enhancements have been made, all arranged by Istat together with

MOS representatives, improving the usability and ease of use, as well as the overall functionality, that overtakes the capabilities of the previous version.

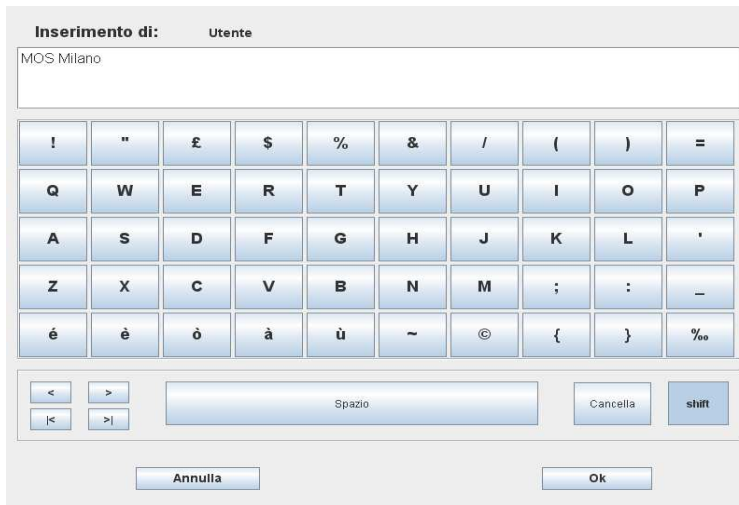


Figure 3: P1 UI - virtual keyboard

are checked against the local DB. Since the hardware equipment has a touch screen but also a small keyboard, the user can choose whether to use the latter or open a virtual keyboard designed on purpose for the need to speed up and ease the insertion of text and numbers. (Figure 3)

At startup P1 performs a self-update: the availability of updates is checked against the centralized server. In case a new version of the software is available, it is automatically downloaded and installed without either user or administrator intervention. This is done by means of SUN's Java Web Start technology.

Once the client is ready to use, it prompts the user for login and password which



Figure 4: P1 user interface - list of outlets

Logged collector can now begin her work. She is presented a frame where she can select between three main activities: a) download from the MOS server the list of quotes assigned to her for the current month, b) start performing the collection of quotes in the assigned outlets and c) send to the MOS all the data collected so far. The first task is generally executed once per month and will setup the local copy of the DB

with the quotes that were collected the month before, so that in case there is no change in the quote the user will just go through a one-click action. On the other hand the upload phase can be executed many times during the collection term.

The core task is the collection itself (Figure 5), which can be started and completed at any time during the term as the quotes have a 'state' in the DB which reflects whether they have been collected or not. At first, the collector is presented a list of outlets along with their address and the number of quotes she is due to collect. P1 clearly

indicates for each outlet if the gathering of quotes has been completed (green highlight) or not (in red, Figure 4).

Figure 5: P1 user interface - collection form

For any given outlet the collector will be able to select the products in the order she prefers and register the new quote. Besides the price of the product, she can use the client to indicate whether there is any change with the quote, either in the quantity, or for the presence of a special discount, or in its peculiarities (brand, type and so on). For any of these specific changes P1 shows appropriate input forms. An indicator shows the

percent variation of the quote so that any typing error can be double-checked by the collector. It is important to notice that it is possible to indicate also the absence of quote for that month.

Before being stored in the database, the collected the quote is also checked for compliance to a set of predefined rules established by the Istat according to the type of product, with the aim of preventing common collection mistakes to occur.

This, as well as the several visual aids included in the interface, should contribute to largely reduce the number of mistakes in collected data with respect to both paper questionnaires and the COBOL software, then significantly enhancing the overall quality of data.

3.4 The Control and Correction Application

The P2 application is a web application also based on Java technology. The choice of a web architecture for such an application was straightforward, since it has to be accessed by several remote users with different access profiles, while development and management are centralized.

P2 uses a combination of Java Server Pages (JSP), Java Beans classes and JDBC Driver connectors in order to grant access to the relational Database, which in this specific case is Oracle version 10.1.2. JSP technology, in combination with Javascript, is also used to deal with the appearance of the web page and to check the client input forms. Tomcat 6 is used as web container. Currently the web application is made up of 60 JSP pages. Given the high number of functions to be supported by the software, a modular design is crucial for achieving a piece of software which is easier to develop, to test and to maintain, without remodeling the application. For example, it is planned to include shortly in the application new modules concerning e-learning and conference functions.

The basic feature of P2 is to allow users to browse data in the central DB, performing manual editing. The editing process is structured into two successive phases, first at MOS, then at Istat. Editing is supported by a sophisticated web interface that, by exploiting on-the-edge web technology such as Ajax, reaches the flexibility of a desktop application. Similarly to P1, complex checks on input data have been also implemented in all the forms, aiding operators to detect and avoid possible mistakes.

MOS can also access functions for managing the daily work of the collectors, distributing the work load. Since data collectors download every day their data, MOS can flexibly change the work load assignment upon necessity (for example, if a data collector falls ill). In order to favor a seamless transition from the old COBOL software, P2 also offers tools for importing/exporting data in the format supported by the COBOL software. This function is allowing a progressive introduction of the new architecture in the production process. Finally, sophisticated reporting functions are available to allow user to create their own customized report, for comprehensive data analysis which is tailored to user needs.

4. Expected Data Quality Improvements

Benefits expected coming from the new IT architecture for CP survey are multiple and they can be resumed in the following ones:

1. the new IT architecture improves the efficiency of the data collection and recovery process (abandoning paper questionnaires, generalized adoption of P1 for data collection and immediate transmission to Istat server, etc.).
2. The availability of one server, on which all the subjects of the production process operate, avoiding the redundancy of data among Istat and MOS and therefore elimination of possible mistakes, harmonizing the IT environment for both Istat and MOS.
3. In terms of quality, accuracy, punctuality and timeliness are expected to improve.
4. The integration of the system allows the development of different set of indicators that drive the activity of MOS and Istat and allow the improvement of data quality.
5. Saving of resources deriving from each statistical progress listed above. In particular the first huge saving of resources comes directly from definitively abandoning paper questionnaires, because data collectors will be able to collect consumer prices more quickly and because office data entry activities will cease. It is measurable in terms of hours worked and resources that could be dedicated to monitoring data quality. Then resources will be saved in terms of hours worked dedicated to keep aligned Istat and MOS archives. Finally resources are saved for the more friendly environment that will allow to standardize and to repeat data queries in a very easy way starting from quality indicators. The amount of resources saved and potentially available for other activities is on the way to be quantified.

Different set of indicators are available for different frequency:

- a) daily indicators that focus the attention on the results of each data collection tour carried out by each collector (amount and percentage of non responses, amount and percentages of replacements, amount and percentages of price changes, time of collection for each elementary observation, outliers, ecc). These indicators allow to promptly correct possible mistake returning in the field, focusing the attention on possible irregular behavior of the collectors;
- b) monthly indicators that focus the attention on the entire data set available for the month and that can highlight possible critical aspects such as too low percentages of price changes or too high percentages of non responses by cause. This set of indicators is based on a system of thresholds that, if overtaken, produce predefined consequences in terms of further checks or estimation or other intervention;
- c) quarterly and annual indicators, mainly dedicated to investigate issues concerning the evolution of the sample and possible inadequacy of the resources involved in data production process.

5. Conclusions and Future Work

In this paper we have presented the new IT architecture supporting the Italian CP survey, currently being developed at Istat. The new architecture will be progressively introduced in the production process. At present, the transition to using P2 is in progress and it is scheduled to be finalized in 2009 March-April. In 2008 November and December, training sessions have been carried out for all MOS that will use P2. Remaining parts of the project (P1 developments and implementation in the field, training sessions for data collectors, definitive implementation in data production process of quality indicators of data and data collectors activity) are expected to be definitively completed by the end of 2009. Therefore the next change of the calculation base (2009 December) of Italian CP indices will be carried out completely in the new IT architecture: that step will be the end of the present transition phase and the definitive passage to new IT environment. This achievement will represent a very relevant step ahead for the quality of CP survey in Italy and of European harmonized data in the field of CPI.

References

- Eurostat (2001), Compendium of HICP reference documents, *Eurostat working documents*
- Gamma E., Helm R., Johnson R., Vlissides J. M. (1995) *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley
- Alonso G., Casati F., Harumi K., Machiraju V. (2004) *Web Services. Concepts, Architectures and Applications*, Springer.