A systematic approach to cleaning fiscal data in Local Public Administration

MARIO BOCHICCHIO, ANTONELLA LONGO
SetLab – Department of Innovation Engineering
University of Lecce
Via per Arnesano, Lecce
ITALY

Abstract: - In the paper we present a systematic approach in managing fiscal data in Local Public Administration. The novel way has been developed to deal with the “avalanche effect”, due to the mismatch of fiscal data in adjacent years, and other similar problems, mainly related to data dirtiness and to the lack of standardization in fiscal data. The application of the approach is discussed referring to the preliminary results of the Taviano Project, a real case in the Southeast of Italy

Key-Words: - Data cleaning, Public Administration, Fiscal Data, Software Maintenance, Data quality

1 Introduction

In Italy Local Public Administrations (PALs) are entitled to manage local taxes like ICI (municipal tax on real estate) and TaRSU (municipal tax on urban rubbish collection). In the last years the whole management process of these taxes, from billing to collecting, was usually performed by proxy for large private companies. In general this produced poor economic results for local public administrations, especially in the smallest towns, so that a number of municipalities are considering retaking the full control of local taxes. This decision requires a substantial effort, both at technical level and at organizational level; the problem indeed is rather complex, and it cannot be solved by simply buying some software applications or some more computers. From the information point of view, for example, needed data comes from a number of different subjects (SOGEI, ANCI, ENEL, Land Register Office, etc.), but no interchange standards have been defined. Moreover, existing data are affected from a number of errors (omission, transcription, communication, etc.), both intentional and/or unintentional. From the organizational point of view, PAL’s officers are often inadequate or insufficient to support this new and crucial process.

In that context, in September 2001, the Software Engineering & Telemedia Lab of the University of Lecce (SET Lab in the following of the paper) in partnership with Servizi Locali SpA (SL in the following), started the Taviano Project, to support a group of local administration in the Southeast of Italy to reorganize the whole local taxes management process. The starting point has been the management of fiscal data in Municipality of Taviano.

The paper describes the main items about the proposed approach and its application. Section 2 gives position of our work compared to the literature. The problem is presented in section 3. The proposed approach is described in section 4. Some preliminary results about the Taviano Project are given in section 5. Conclusion and future works are in section 6. Section 7 is for bibliography.

2 Related works

Providing citizens with e-services is a hot research topic today. The development of new interaction paradigms and the use of communication networks and of distributed applications to offer new added value services to citizens have been undertaking in many countries ([B4], [B5]), under the wide umbrella of what is referred to as e-government. Examples are the European project “eGOV” ([B1],[B2]), the Italian Public Administration Network ([B3],[B6]), the experiences about the Municipal Transformation in Norway ([B7]).

These experiences focus on the implementation of a single point of access to public services and information, the development of integrated platforms, common Cooperative Architectures, standardized data and communication exchanges, which will allow the public sector to provide citizens, businesses and other public authorities with information and public services structures. The mentioned services and structures lay over the integration of information flows coming from different public agencies; interesting contributions show how to integrate heterogeneous legacy information systems ([B8], [B9]) in PA, nevertheless these services assume source data are of good quality. Data quality can be defined [B10] as the measure of the agreement between the data views presented by an information system and that same data in the real world and Data Quality Management (DQM) is the set of procedures necessary to manage and maintain data quality. Ballou and Pazer [B11] identified and discussed four dimensions of data quality: accuracy, completeness, consistency, and timeliness. Accuracy could refer to recording correctly facts, completeness to having all relevant information recorded, consistency to a uniform format for recording the relevant information, and timeliness to recording the information shortly after the disposition. Data quality concerns arise when these dimensions are not verified in a given context and this often happens when one wants to correct anomalies in a single data source (e.g., due to misspellings during data entry, missing information or other invalid data), or when one wants to integrate data coming from multiple sources into a single new data source (e.g., in data warehouses, federated database systems or global web-based information systems). *Data cleaning*, also called *data cleansing* or *scrubbing*, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [B12] (e.g. accurate and consistent data, consolidation of different data representations and elimination of duplicate information). On the market a large variety of tools is available to support data cleaning tasks. The philosophy behind [B13] is either inference based (i.e. cleaning is performed discovering patterns, within the data and using these patterns to discover rule to data clean) or data based (i.e. concentration on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination). Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of cleaning problems. Our experience shows why management of fiscal data is far away to be easily handled today; later in this contribution we will demonstrate the four dimensions of data quality are not respected and convenient frameworks supporting efficient data cleaning process in PA have being just starting to be designed.

3 Problem Formulation

Taviano is a Municipality in the Southeast part of Italy with 11.000 inhabitants. Two years ago (in 2000) the Municipal authority decided to transform Taviano in a laboratory of innovative interactive services to government users (both outside the government – citizens, institutions, business – and inside – employees): what we call *Citizen Relationship Management*. The first step was to design and support an effective tax management process (the *Virtual Office of Incomes*), in order to give the Municipality the chance to know its actual incomes and to plan the budget on "certified" data. Beside regulations, organization and process to be changed to support an efficient and effective service to users (both outside and inside the Municipality), the first issues we run into were data quality assessment and the development of a systematic approach to clean data, since information we needed to deploy the Virtual Office of Incomes came from data archived on paper and unreliable data flows from other institutions or from legacy applications

To understand the reason behind our approach, it's necessary to say about the Italian taxes management process. A good example, concerning the management of ICI (the Municipal Tax on Estates, established in 1993, about the ownership of real estates), is shortly described in the following subsection.

3.1 ICI Management process

For several years citizen have had to calculate the amount and pay it through the concession private agency. The private dealer had to collect payments and eventually to sample audit them with data coming from the Land Registry Office in order to assess possible tax evasion. The main reasons why Public Administration doesn't calculate bills have been:

- ❑ The basic approach to taxpayer, almost "oppressed" by Government, who had to undergo requirements, norms and twisted bureaucratic rules, often difficult to be understood by common people, who had to be responsible to know about duties and to accomplish them.
- ❑ The consciousness of the dirtiness and unreliability of data owned by Public Agencies, very often maintained and archived on paper, and the consequent impossibility to calculate the exact amount of the tax in reasonable time and with reasonable effort for all citizens

Each year citizens must pay ICI according to a formula depending on the percentage of ownership, the land registry revenue, the ownership of other estates, and some more annual parameters.

If any variation about the estate occurs during the year, the taxpayer must send a statement to the Municipal Tax Office about the change of the fiscal parameters and he must pay according to the new parameters. If no variation occurs during the year, the tax is calculated using the previews year's parameters.

As said before fiscal data are rather difficult to collect and to manage because of the lack of standards in data format and in the related semantics and because of the unreliability of existing data

Over the years we verified Local Government changed the way the tax has been calculated: in 1993 the percentage of ownership was not considered, and it was added the following year, but nobody asked to taxpayers to integrate the missing information. The result is that assessing payments in a year and crossing data from different sources with reference to 1993 payments, two people sharing a house either declared the same estate twice or mismatched the information from the local land registry office and from the ICI statement. This fact can be a reason for fiscal assessment from the Municipality.

Data unreliability also comes from type writing mistakes, misspellings, missing data (like SSN, the address, etc.), data conversion from different formats and different representation of the same information (like "via Reg. Margherita" that is the same as "Via Regina Margherita" and so on) causing a wrong recombination during cross assessment among different offices and information sources. So, for a given taxpayer, data coming from various years can be not correctly recombined; likewise, data coming from different taxpayers can be erroneously put together

The reference to the previous year is the reason why if an error occurs in year X, it will happen in year X+1 as well as in the following years, and this is the main reason of the "avalanche effect" we observed systematically in our fiscal database. We mean that after the declaration was made the first year, citizens has been asked to communicate to the administration only the variations of fiscal data with respect of those previously declared. The propagation of the various errors to the subsequent years is devastating for the correctness of the fiscal calculations: for example, for the Municipality of Taviano (about 6000 taxpayers), starting data collection from '93, fiscal records related to '98 are wrong at 50% (one record out of two is wrong !!!) and this is what we call "avalanche effect".

4 The Yes approach

The approach we developed, aims to limit as much as possible the avalanche effect. It can be divided into 3 main actions:

1. *Data Correction*: concerning the finding and correction, when possible, of domain inconsistency, out of range values, absence of compulsory values, etc
2. *Data Normalization*: concerning the validation of data against a given set of values and/or rules;
3. *Year Separation*: concerning the integration of the various information sources needed to perform the ICI's calculations and the related evasion detection. This is the most important step to limit the avalanche effect, and it is based on the logical separation of the information coming from the various years.

We can observe that the first two steps, dealing with single source cleaning issues, are interesting topics in the research community, as also shown in section 2, while the last step, concerning the integration and the linking of multiple information sources, is peculiar to the proposed approach.

4.1 Data Correction & Data Normalization

All the archives needed to ICI's calculations and evasion detection were subjected to this actions, with particular attention to data from SOGEI (the firm collecting ICI's statements in 1993), coming in textual format with a lot of lacks or misspellings in SSN, Cadastral numbers, citizen names and addresses. As also stated also in [B12], for sources without schema, such as files, there are few restrictions on what data can be entered and stored, giving rise to a high probability of errors and inconsistencies. The cleaning phase concerned the correction, when possible, of errors, like domain inconsistency, out of range values, absence of compulsory values (to be retrieved from other information sources).

Usual problems we run into have been ([B12,][B13]):

- ❑ *Values embedded within free forms values need to be split out into separate attributes*: in 1993 statement data name and surname were embedded within the name.
- ❑ *Misfielded*: the concatenation of name and surname put into the SSN field
- ❑ *Noise data*[1]: the Cadastral Id is a number, but we found some instances as "Z0000"
- ❑ *Missing data*: missing SSNs, addresses, etc
- ❑ *Duplicates:* in 1993 ICI statements were identified by the Cadastral number and the SSN, from 1994 on they were identified by the SSN and the combination of 3 fields substituting the Catastral number, moreover every source uses different its own identifier for citizen and estate

The normalization phase concerned the "quantization" of the dictionary from the actual archives on a "normal dictionary" from more reliable historical archives of Taviano and/or the application of functional rules over the existing data. Dirty data types we were able to normalize have been [B13]:

- ❑ *Misspellings generating errors*: Errors in names, surnames: "AnnMaria" instead of "Anna Maria".
- ❑ *Strays from business rules*: sometimes SSN doesn't match the calculation made on personal data
- ❑ *Strays from business rules*: sometimes SSN doesn't match the calculation made on personal data
- ❑ *Varying representation*: the sex is represented like "0/1", "M/F".
- ❑ *Varying semantics*: the Cadastral number, which is the unique identifier in 1993 ICI statements, is a unique field in SOGEI flow; in years 1994 and following the Cadastral Number is split out into 3 fields.

4.2 Yes Separation

The "YeS" (acronym of Year Separation) approach has been developed to limit and to put over control the negative effects of the "avalanche effect". It is based on a strict separation, both at logical level and at physical level, of the fiscal data coming from the various years.

As shown in Fig.1, for a given a year, to detect potential tax evaders (Phase 1), we look for discrepancies among data coming from Paying – in slips (corresponding to the versed amount), from the due amounts (calculated from the Land Registry Office archive and from the ENEL archive) and from the stated amount (from ICI statements). A Tax Assessment Request is sent to citizens (Phase 2), who can (Phase 3) either accept their position of total or partial evaders or prove the mistake and request the related correction. If the taxpayer must pay (Phase 4) he can either pay off the sanction or he will receive an injunction to pay

[1] Values outside the domain attribute

at the fixed deadline. The last phase can be repeated until the taxpayer position clears up

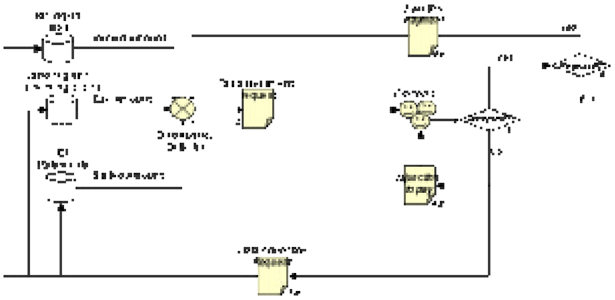

Fig. 1. Yes approach: the logic model for a given year

Communications (Tax Request Assessment, Sanction Payment, Data Correction Request) between citizens and Local Administration are all stored in the "*Communication Archive*" (not shown in Fig,1 to keep the picture cleaner) to keep track of the interaction between Citizen and PA. This archive is another input to the Discrepancy Detector, in order to put right the position of the taxpayer, avoiding to sending him other tax assessment request after their paying-off.

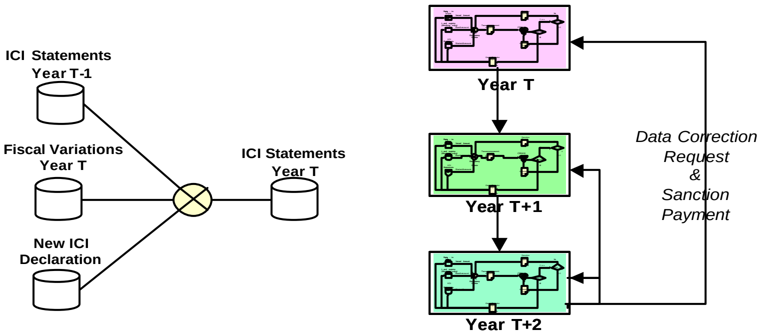

Fig. 2. Links among years: the ICI statement archive and Data Correction Request and Sanction Payment Communications.

The approach to assess ICI is the same every year, but usually tax auditing is delayed of some years, so the YES approach needs to consider the relationships among years.

We have located relationships among years in two points: the ICI statement archive for a given year and the communications flows of the Sanction Payment and the Data Correction Request. In a given year ICI statement is made of the union of ICI declaration sent during the given year, and the fiscal variations over the previous year's statements presented during the year. This means that a correction and/or error in a year affect all the following ICI statements.

If the fiscal audit is made in year T+2 for year T, the Sanction Payment and the Data Correction Request communications are referred to Year T and also impact data in year T+1.

The Discrepancy Detector integrates multiple data sources. Usual issues we met with have been [] naming conflicts, structural conflicts, different representations or interpretations of values, different aggregation levels or reference points, duplicates contradictory values.

5 Project results

In Fig.3 an Fig. 4 the diagrams showing the avalanche effect at the begin and after the application of the 3 phases of the YeS approach are represented for the town of Taviano, from '93 to '98.

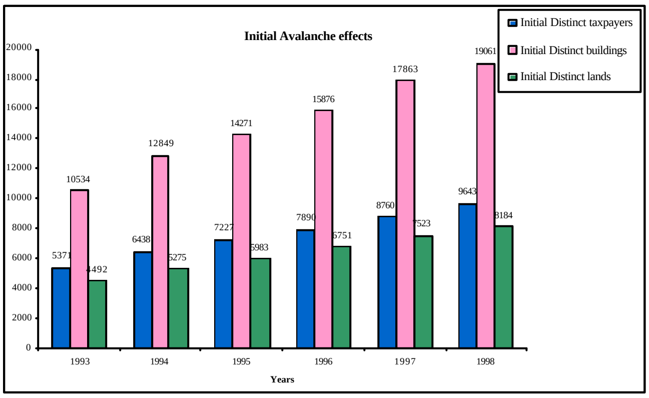

Fig. 3. Avalanche effect before the Yes approach

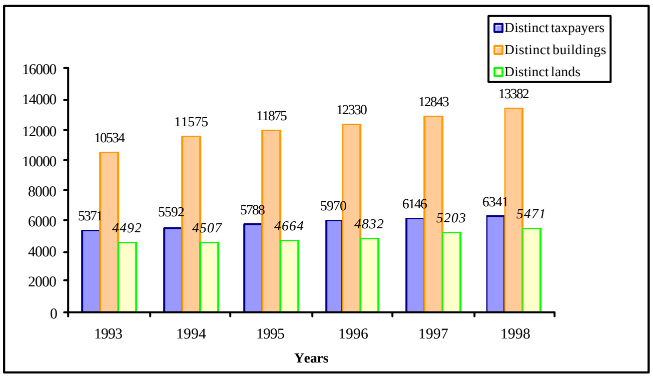

Fig. 4. Avalanche effect after the Yes Approach

In comparison with the 50% of wrong record previously reported for '98 (uncorrected avalanche effect) we can state that the YeS approach is very useful to effectively reduce the "wrong evasion detection" without reducing the "true evasion detection". After data cleaning with the yes approach we obtained the increase of the number of about 18% from 1993 to 1998. Fig. 5 shows the shift in the "avalanche effects" before and after treating data with the Yes approach. Municipal employees state that as a rule of thumb the number of taxpayers increased of 45% from 1993 to1998.

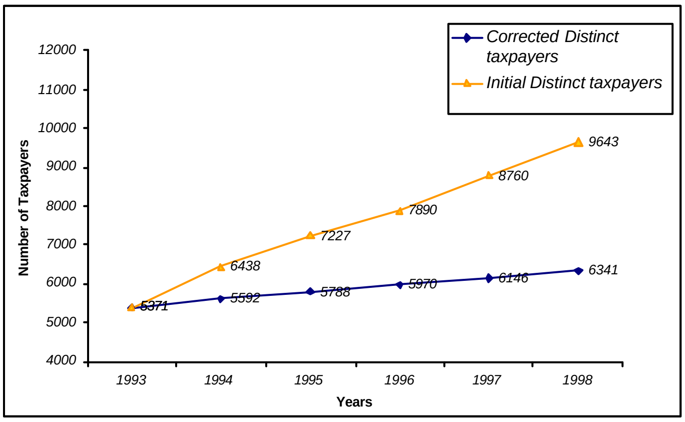

Fig. 5. An example of the shift before and after the data cleaning

From the Local Government point of view the cutback in number of errors means:

- reduction of disservice in tax management they have been fighting for years because dealers managing taxes usually crossed data without any cleaning, sending tax assessments to a taxpayer every two
- less costs in managing assessment with paper sent, less-front end time to answer to angry people
- better consciousness of the actual incomes and to better plan the new year's budget.
- gain of trust, because only people with irregularities or odd data are detected

6. Project results

In the paper we described a 3-steps approach to effectively perform tax calculations for Local Public Administrations.

The approach is very robust with respect to various errors types, which are no more propagated to the successive years, and this is extremely important to reduce the number of "false detections" of tax evaders in the system (avalanche effect).

The proposed approach has been verified in a real project, with excellent results.

It is rather general and scalable, so it can be easily extended to large communities of citizens.

About 15 municipalities, with more than 200.000 inhabitants recently delivered to use that system for their local tax management activities.

A systematic evaluation of the achieved results will be performed to better estimate the effectiveness of the approach and to decide the future steps, both at theoretical and at practical level.

References:

[B1]. Wimmer M, Traunmueller R.: Towards an Integrated Platform for Online One-Stop Government, in: ERCIM News No 48, Special Theme: e-Government, January 2002, pp. 14-15 (ISSN: 0926-4981)

[B2]. Traunmüller R., Wimmer M.A.: Web Semantics in e-Government: A Tour d'Horizon on Essential Features. In Proceedings of the WISE'01 Workshop on Web Semantics, Kyoto, December 3-7, 2001

[B3]. Batini C., Mecella M., Cooperative Processes for e-Government, in: ERCIM News No 48, Special Theme: e-Government, January 2002, pp. 38-39 (ISSN: 0926-4981)

[B4]. Central IT Unit (CITU) of the Cabinet Of.ce. Information Age Government. Benchmarking Electronic Service Delivery. CITU Report, London, July 2000

[B5]. Elmagarmid A.K., McIver Jr, W.J. (eds.). The Ongoing March TowardDigital Government (Special Issue on Digital Government). IEEE Computer, 34:2, 2001

[B6]. Mecella M., Batini C.: Enabling Italian E-Government through a Cooperative Architecture in A.K. Elmagarmid, W.J. McIver Jr. (eds.): Special Issue on Digital Government. *IEEE Computer*, vol. 34, no. 2, February 2001.

[B7]. Beck E.: E-Government and Municipal Transformation: Some Experiences, in: ERCIM News No 48, Special Theme: e-Government, January 2002, pp. 21-22 (ISSN: 0926-4981)

[B8]. Mecella M., Pernici B.: Designing Wrapper Components for E-Services in Integrating Heterogeneous Systems. VLDB Journal, vol. 10, no. 1, 2001. Springer-Verlag

[B9]. Pernici B, Mecella M, Batini C.: Conceptual Modeling and Software Components Reuse: Towards the Unification in A. Sølvberg, S. Brinkkemper, E. Lindencrona (eds.): Information Systems Engineering: State of the Art and Research Themes. Springer Verlag, London, 2000. Springer-Verlag)

[B10]. Orr K.: Data quality and Systems in: COMMUNICATIONS OF THE ACM, Vol. 41, No. 2, 1998, 66-71

[B11]. Ballou, D.P. and Pazer, H.L. Modeling data and process quality in multiinput, multi-output information systems. Management Science 31, 2 (1985), 150–162.

[B12]. Rahm E., Do H. H.: Data Cleaning: Problems and Current Approaches in Bulletin of the Technical Committee on Data Engineerind, Special on Data Cleaning, December 2000 Vol. 23 No. 4, pp. 3-11

[B13]. Quass, D.: A Framework for Research in Data Cleaning. Unpublished Manuscript. Brigham Young Univ., 1999.