Paper 122-2010

ELISA: DATA QUALITY IMPLEMENTATION TO TAX EVASION (A LOCAL PUBLIC ADMINISTRATION EXAMPLE)

Paola Leproni, Data and Project Coordination, CSI-Piemonte

ABSTRACT

Since 1977, CSI-Piemonte has promoted innovation in local public administration through the use of the most modern information technology and Internet tools. Thanks to its contributions, the Piedmont Region operates on the Italian and international scene as an integrated administrative system equipped with the necessary technological infrastructure. With nine offices operating in the region and 54 consortium members, CSI is one of the largest Italian ICT operators in the public sector. This presence on the territory simplifies the communications with organizations, citizens and enterprises and enables it to offer simple, efficient services contributing to re-launch the regional economy.

In 2003, CSI-Piemonte decided to implement a Business Intelligence Competency Centre in SAS® 9. The BICC grew over the years and today consists of 78 people operating in different areas:

- Metadata management, data re-usage standards (1471 databases managed)
- Decision support systems design and development
- Decision support systems for HR management, strategic control management
- Catalogues, cross-area data, production activities observatories
- Health care decision support systems
- Data quality and Data, Text Miner

The aim of CSI-Piemonte is to capitalize the strong and longstanding experience in SAS on new international projects and to offer it to support the developments of new applications for the Public Administration in foreign countries.

CSI-Piemonte was awarded the SAS “Enterprise Intelligence Award” in Stockholm during the SAS Forum 2007 for decades of experience gained in the field of business intelligence.

In this presentation we put in evidence our experience on Data Quality and we’ll describe the **ELISA Program** (System Innovation for Local Authorities). This project, realized for the City of Turin, is the integration in an Operational Data Store of all data belonging to local tax management systems, registries’ official data sources and cadastres. We obtain this integration by the construction of data cleansing services and decisional systems in order to allow all local authorities to appropriately supervise real estate movements within their territory.

THE COMPANY

CSI-Piemonte (Consortium for Information Systems) was founded by the Piedmont Region, the University of Turin and the Polytechnic of Turin in 1977 with the aim of promoting the modernization of local administration by using the most advanced information and IT-based tools to create information services and systems.

As a common meeting point for research bodies, local Public Administration (PA) and private business, CSI-Piemonte looks to spread the benefits of the Information Society throughout the territory and encourage socio-economic growth in the Piedmont region. Thanks to the consortium, Piedmont has become an integrated administration “system”, able to face the e-government challenge, simplify administrative processes and meet the expectations of citizens and companies alike.

Well-known both at national and international level, CSI-Piemonte represents a unique player in terms of it being a 'public' consortium organized along 'private' lines. Nowadays, it is one of the largest Italian ICT operators in the public sector.

Furthermore, the company abides to "socially-responsible" standards of behaviour, maintaining close contact with other key bodies and institutions in Piedmont and meeting the demand for innovation by taking into consideration the natural, social and human resources within the territory.

Specific strategic objectives guide the consortium's actions: encourage the development of the Piedmont System (SistemaPiemonte); promote PA employee training and development; reorganize the range of services available to the Piedmont health department; develop the Piedmont Broadband Network Plan; favour research and innovation in order to support the economic-productive system in Piedmont; carry out the second phase of e-government, encouraging dialogue between local administrations and central government, and, finally, develop its skills and ability to act internationally by working alongside developing countries and being actively involved in European research and development projects.

CSI-Piemonte operates in several different fields of activity, which can be summarized as follows: agriculture and forestry, the environment and territory, demography, land register and taxation, production activities, vocational training and work, education and the cultural heritage, healthcare and social welfare services, administration, accounting and personnel systems.

Key Figures

Consortium members: 80 (including municipalities, municipality associations, local health bodies, hospitals, etc.)

Annual turnover: more than 175M € (in 2009).

Number of office locations in Piedmont: 6 (and one branch in Brussels).

Number of personnel: over 1,200 employees

INTRODUCTION

CSI-Piemonte is the regional Public Consortium with public right legal entity status that operates in the Piedmont Region to:

- design, develop and manage the stakeholders Information Systems and for this purpose it is the recipient of the powers of the Authority for IT in Public Administration;
- promote and create continuous collaboration modes between territorial Institutions and Universities in the fields of:
    - research and development of new IT technologies;
    - their transfer to services both in the PA and production structures;
    - training aimed at these technologies or mediated through them;
- set up an organization and technical pole of the Public Administrations located in the region, to interconnect them on a provincial, local or municipal basis, in compliance with the directives of the Authority for IT in Public Administration.

The role of the Consortium is that of a key instrument for Piedmont PA reform, through the interaction between the public information systems on the Public Administration Network (RuparPiemonte) and its mission is the setting up of the "Piedmont System" for the implementation of administrative decentralization using ICTs.

Our "Data and Projects Coordination Directorate" focuses on data with the following objectives:

- designing and setting up Data Warehouse and Business Intelligence services;
- favouring the normalization of existing data banks, favouring design and production of new integrated data banks and the reuse, with particular attention to their availability and usability in a network;
- contributing to the identification of reference technologies on these topics.

The catalogue of the data and services provided for the members of the consortium makes it possible to draw a picture of the existing data bases.

	TYPE OF DB		TOTAL
PA CUSTOMER	DECISION MAKING	OPERATIONAL	
Piedmont Region	108	854	962
Municipality of Turin	33	197	230
Province of Turin	6	123	129
Others	13	137	150
TOTAL	160	1.311	1.471

(data updated in September 2009)

Many of these data bases contain sets of information that can be used for analyses and consultations (that is Data Warehouse or Data Mart).

These data bases provided the foundations for different types of analysis and access tools, that can for the greater part be accessed in web mode. The following table summarizes the number of services set up according to typology:

	Front-end							Infrastructure services	Back-end		
PA CUSTOMER	Q&R BO	Q&R SAS	OLAP (only SAS)	Dashboard SAS	Dashboard BO	Text e Data Mining (SAS)	Catalogue		traditional ETL	DQ	Total
Piedmont Region	40	54	22	2	1	2	1		84	22	228
Municipality of Turin	15	25	5	1		1	1		39	6	93
Province of Turin	1	3	2	1			1		4		12
Heath (Piedmont Regione)	5	6	2				1		11		25
Local Health Agencies/Hospitals	2	2		2			1		1	1	9
Cross customers	2							1	1		4
CSI	1	10		1			1		15		28
TOTAL	66	100	31	7	1	3	6	1	155	29	399

(data updated in September 2009)

Key:
Q&R BO = system that makes it possible to produce reports, set up with Business Objects technology
Q&R SAS = system that makes it possible to produce reports, set up with SAS technology
OLAP (only SAS) = system that makes it possible to carry out multidimensional analyses, set up with SAS technology
Text e Data mining (SAS) = advanced statistical analysis experiences through the application of data mining (extraction of useful information, carried out in automatic or semiautomatic mode, from large quantities of data) or text mining (extraction of useful information from large quantities of written text), set up with SAS technology
DQ = applications that use data cleansing technologies (in massive mode and with direct interaction on operational systems)
Dashboard SAS e Dashboard BO = system that makes it possible to represent in short (through a dashboard) a series of summary indicators for a specific phenomenon
Traditional ETL = periodic update processes for data bases, that apply integration techniques

BUSINESS INTELLIGENCE COMPETENCY CENTRE

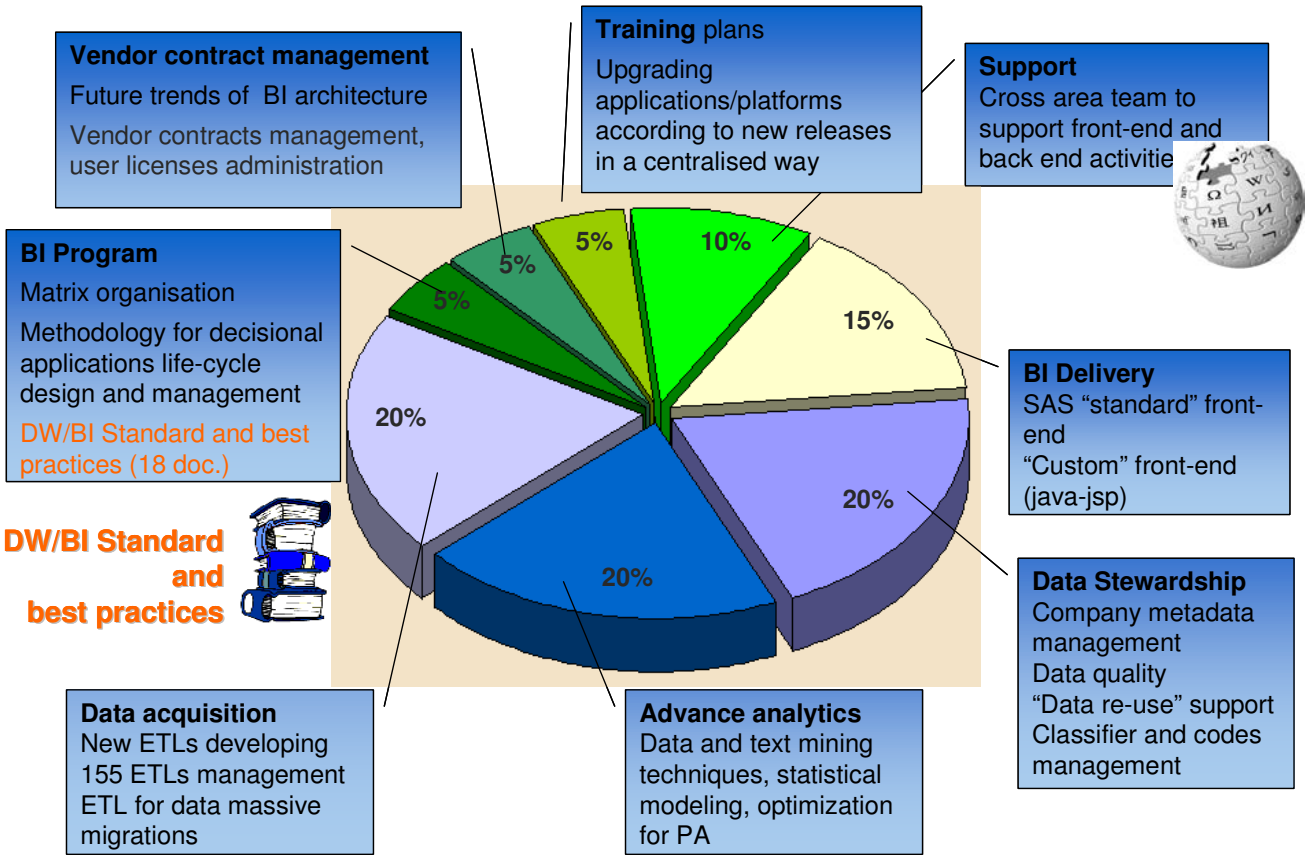

Figure 1 - BICC

CSI created a Business Intelligence Competency Centre (BICC), to support its clients, which provides specialist support by telephone or in person through a centralized call-centre for all the Piedmont Region. Competences are developed in those technologies: SAS®9 and Business Objects.

The Business Intelligence Competence Centre has a matrix organization and it consists of 78 people (including both consultants and employees) whose skills can be resumed in the following activities:

- **Data stewardship**: specialization in metadata management and governance (both from the business and technical point of view), experience on massive and accurate data cleansing and data quality, data reuse support, cross sectorial decoding table classification and management.
- **Data Acquisition**: use of Extraction Transformation and Loading process to develop new decisional processes, to manage services already in production (155 ETLs) and for massive databases migrations.
- **Advanced Analytics**: specific knowledge in Data and Text mining techniques, in the definition of statistical models and in the public administration optimisation processes.
- **BI Delivery**: SAS and BO vendors offer basic tools for the diffusion of BI applications, but CSI-Piemonte uses more customised solutions that are realized (with java and jsp) according to the specific requirements of the customers of the Piedmont Public Administration.
- **Business Intelligence Program**: the matrix organization offers a centre of excellence for the design and governance of the whole life cycle of decisional applications. Indeed, the BICC has defined some

methodological standards and specific polices to support this process.

- **Vendor Contracts Management**: a specific group deals with the evolution of the Business Intelligence technical infrastructure, also dealing with the contracts for administration licences.
- **Support**: a cross-department group was defined to support back-end and front-end activities, to have a centralised area for the collection of the issues and for a first level solution. If necessary, this group refers to the technical support of the relevant vendors.
- **Training**: management of training plans to spread internally the competence on existing BI tools and the new forms introduced in the platform. A very high technology knowledge level is always maintained.

Our competence centre has organized direct training opportunities for ICT companies of the Piedmont Region from 2006, so that companies can train and maintain pool of resources specialized in SAS V9 (MasterClass SAS) and BO technologies. This opportunity helps companies to exploit possible initiatives in the international market.

DECISIONAL SYSTEMS

Business Intelligence systems create a true information value chain, which grows and evolves into an essential preliminary element for decision-making at all levels (i.e. political, managing, operative). Thus, data are transformed in information, knowledge and, finally, decisions.

Our customers (Piedmont Region, Provinces and Municipalities) have different roles in territory administration:

- strategic (Piedmont Region)
- tactic (Provinces)
- operative (Municipalities)

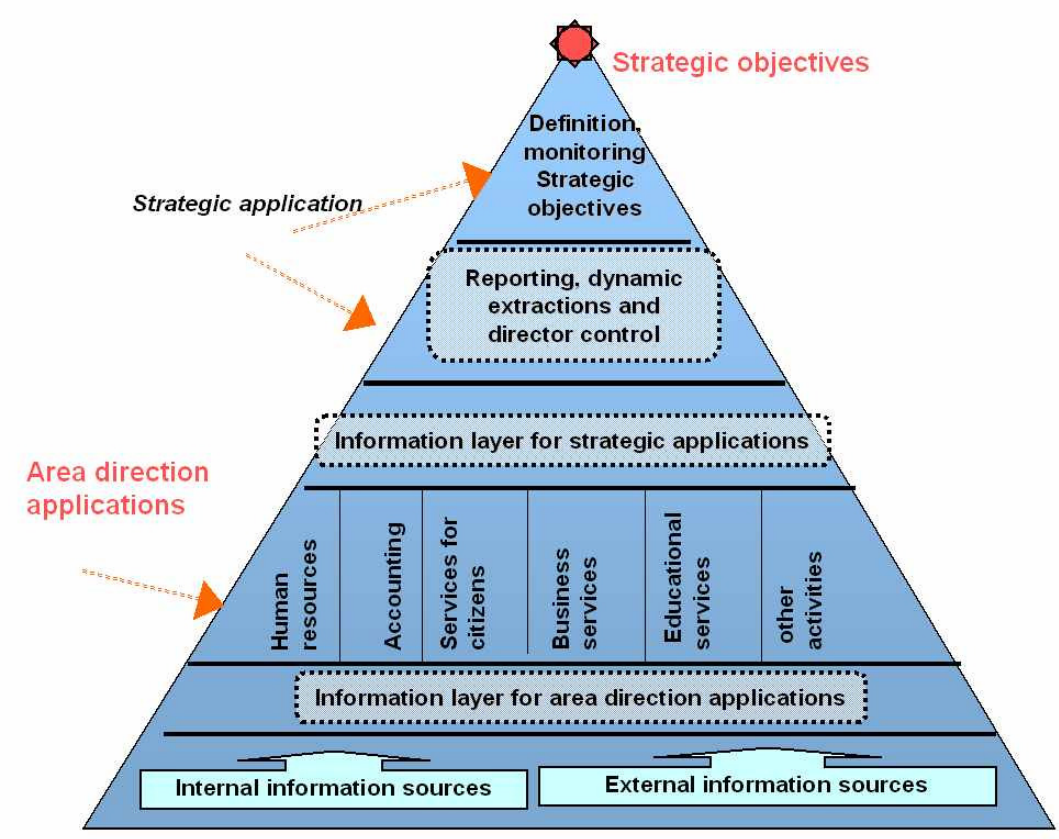

Figure 2

A Decisional Service Pyramid (Figure 2) that provides different kinds of service was set up to support those different roles:

- programming and planning functions for the Regional strategic level
- control functions for the Provincial tactic level
- citizens and companies administration functions for the Municipalities operative level

However, the different customers make requests for the other two typologies as well, so at times CSI provides strategic services for the more operative institutions.

The applications were in a first phase realized in a client/server architecture, while at a later stage, because of the large number of users, they have been migrated to a mostly web architecture, according to the Centred Service Provider logic.

The Business Intelligence platform is completely integrated in the CSI Server Farm and it was designed to provide a single and centralised platform for all the possible customers of the company.

Figure 3 highlights the interaction and cooperation mechanisms of the decisional systems towards the management systems through the so-called "intelligent agents" (Enfine/alert engine layer), that is on the basis of information located in the data warehouse it is possible to set off mechanisms that interact directly on the operating systems.

The evolution does not consider the decisional systems as isolated from the operating systems, but rather as closely integrated with continuous interaction operations. In addition to the data warehouse (container labelled as DW) there is a data container labelled as RealTime DataStore (RTS container): it is a sub-set of operating data duly fed according to the needs and updated quite frequently. This container is used for all the reporting activities that do not require a denormalised structure and that involve a de-normalised structure and that involve limited data quantities and a very frequent update. A first example of such interaction is the use of Data Quality techniques, on-demand, by the same operating environments.

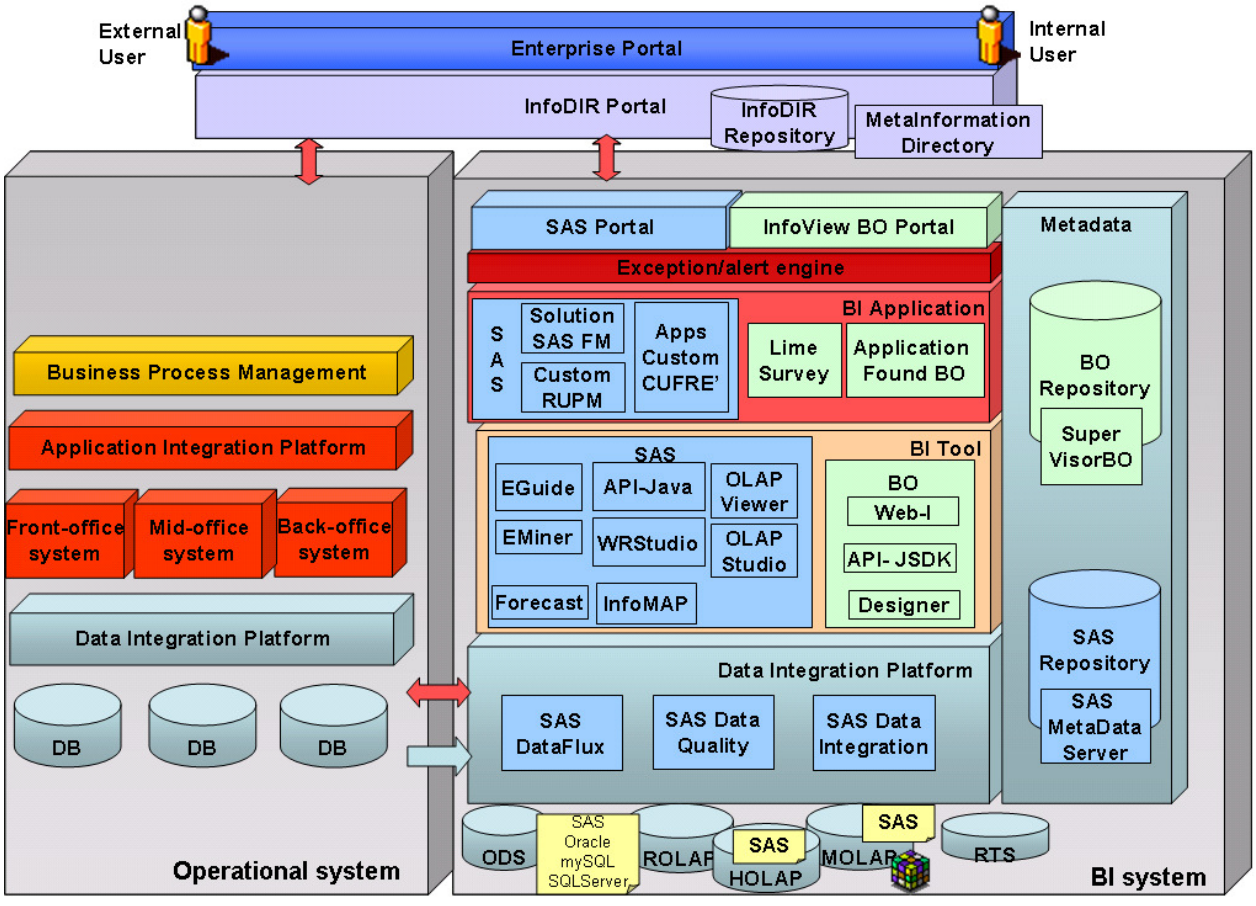

Figure 3 - Technology CSI Framework

DATA QUALITY ACTIVITY

Year 2004 saw the adoption of a Data cleansing specialized tool, as a support in all the phases of the complex data cleansing process, that from the initial analysis of the problem (data and process analysis), continues with the planning and implementation of the improvement interventions and ends incorporating the improvement measures into the system.

Simultaneously to the adoption of the tool, Data Quality activities have begun on data on street directories, compulsory schooling, employment and taxes. The activities undertaken can be grouped along the following typologies:

- Analysis of the quality of data and identification of operating and decisional data base anomalies (quality assessment);
- Data cleansing and standardization interventions;
- Interventions on data acquisition processes to favour the elimination of the problem and to guarantee the reproducibility of the cleansing interventions;
- Data migration with data control/cleansing functionality.

Hereafter is a summary of some of the issues faced up to now, in the framework of the activities undertaken:

- Personal detail control and recognition issue (natural and legal persons): sophisticated data correctness and coherence controls are applied to personal detail records with the identification of multiple records that refer to the same individual (search for similar instances). Moreover, recognition processes are carried out using reference data banks (For ex. Registries)
- Address normalization and recognition: addresses are standardized and modified to fit the required format (for ex. type of street, street name, extension, number), while the address that are not codified are checked against a reference street directory;
- Identification of the place of birth/residence issue: the transversal tables relating to administrative limitations and foreign countries are used with this regard: the objective is that of identifying a place of birth or residence entered in a descriptive way to an official coding or denomination.

From 2004 to today, many data cleansing and porting projects have been implemented or are under implementation, and here are some examples.

Compulsory Schooling Data Cleansing: Project on data cleansing and redesign of the data bases and feeding procedures of the regional data bank on compulsory schooling, containing the personal details of the students of compulsory schooling age.

Municipal Tax Register Porting and Cleansing: Analysis, data cleansing and migration project from the present data bank for the management of municipal tax payers - ATC (on mainframe) – to the new GMS (open) Subject World Management system .

Road Tax Register Porting and Cleansing: Register data analysis, cleansing and migration from the Road Tax regional system to the new regional GMS.

Piedmont Employment Information System (SILP) Data Porting and Cleansing: Data migration and cleansing of the 20 Netlabor and Prolabor employment provincial systems in the new SILP integrated system.

Economic and Production Activity Register Address Normalization Service: the address normalization service ensures cleansing and standardization of the addresses of the enterprises with data quality functions and using the Regional Road Directory as reference data bank.

Post Office Component for Municipality of Turin Applications: transversal service that makes it possible to structure the addresses whose input comes from all applications, in a file that is articulated according to the standards of the Italian Post System for the use of new post services. Moreover, it makes it possible to cleanse and structure the addresses according to the reference toponymy.

AURA (Unified Regional Patient Archive): integration in the new regional centralized register of all data belonging to the 15 local health authorities' and the 8 hospital agencies' databases by means of record matching techniques; application of data cleansing ad data normalization functionalities.

ELISA Program (System Innovation for Local Authorities): integration in an Operational Data Store of all data belonging to local tax management systems, registries' official data sources and cadastres; construction of data cleansing services and decisional systems in order to allow all local authorities to appropriately supervise real estate movements within their territory.

ELISA: DATA QUALITY IMPLEMENTATION TO TAX EVASION

ELISA is a DAR program (Regional Affairs Department, Cabinet Presidency) in order to co-found technology innovation projects developed by local authorities.

All the projects that are involved in this National Program share these general objectives:

- promote digitalizing of administrative activity;
- have a national value;
- ensure the territory growth;
- ensure the replication on the territory;
- be consistent with regional plans

The Local Public Administration involved in this project is the City of Turin, that asked us to focus on two lines of funds: territory and tax. Our objective is the creation of a unified system concerning taxation to integrate all existing solutions, to provide taxation supporting tools, both in terms of analysis and to alert on tax evasion and to provide tools able to support paying citizens, both in terms of consultation of their tax position and of management and execution of payments.

We reached the objective by the implementation of

- a Subject/Object/Relation Municipal Register (ACSOR)
- a local Analysis data warehouse and Dashboard for local taxes' collection
- a Taxation Dashboard for Treasury taxes' verification

The general schema for the project is described in the following figure:

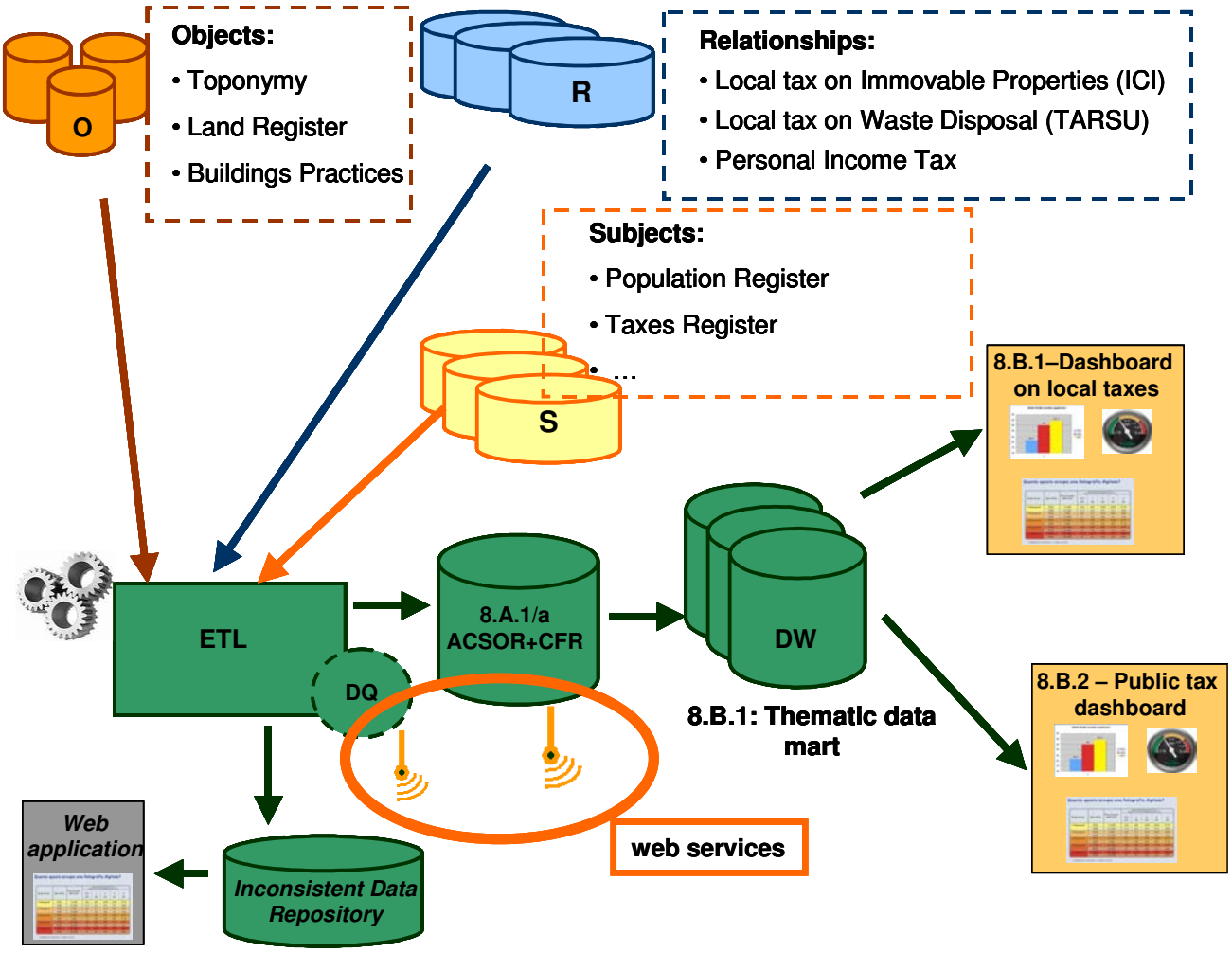

Figure 4 - ELISA Schema

We collect all the data about Objects, Subjects and Relationship that are involved in tax analysis and we use Data Integration and Data Quality to create the Operational Data Store ACSOR.

The ACSOR is the new Municipal Register that contains Subject, Object and Relation:

- Municipal Subject Register (ACS), allows subject reconciliation, by means of unambiguous identification
- Municipal Object Register (ACO), maximizes system's capability of univocally identifying objects
- Municipal Use Relation and Property Register (RUP), defines a standard method for representing all status/relations deriving from occupation or possession of objects located on the municipal territory

The Citizen Facts Repository (CFR) allows correct acquisition in the data warehouse of "non-register" information, that is fines, payments, etc.

The Inconsitent Data Repository stores cleansing and reconciliation choices and will be made available to municipal officials by web.

The Data Quality functionalities are:

- Cleansing and data transformation activities
  - data quality assessment for all sources involved (in order to define their quality level and to suggest some possible data cleansing activities for the future)
  - criteria definition to identify a valid entry in a cluster group (application of a weight-and-measure system that determines a record ranking in a cluster group)
  - information normalization, by means of lexical/syntactical algorithms and reference data bases (addresses)
  - cleansing of incorrect or completing lacking information, inferring them from other fields that are correctly filled in, with formal checks on fields (fiscal code's correctness, coherence of fiscal code with personal data, coherence of gender with first name, etc.) and identification of duplicated data (application of "record matching" techniques, identification of similar record groups)
- Data Reconciliation
  - source schemas' normalization, explicating the "subject", "object" and "relation" entities
  - integration in the reconciled schema, identifying the best-of-breed record by means of tools allowing information comparison both in terms of sameness and similarity

CONCLUSION

One of the Consortium's transversal tasks is to add value to the large quantity of existing data found in the Piedmont Region's archives, as this information is one of the most valuable resources available for improving the region's administrative mechanisms.

The heterogeneous nature of the platforms and formats could threaten the quality of the information and, in short, the decisions that depend on it. For this reason and through its experience in the use and response of SAS technology in many areas of information management and handling, CSI-Piemonte trusts SAS to guarantee the quality of the data.

In the ELISA project we use Data Quality techniques in order to increase the quality of the data source to analyse tax evasion because it's very important to have a pool of more accurate data to save time in searching tax evaders, to obtain full information and high quality comparing different sources because the single sources can be incomplete.

The application of improvement methodologies on data acquisition processes with the ACSOR's web services, provided by the DfIntelliserver module, can incorporate quality control on data, in order to maintain the quality reached by means of the massive intervention over the long term.

The main focus of the project is to build one single framework of rules to share the benefits along all Municipal Officials.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Paola Leproni
Enterprise: CSI-Piemonte
Address: C.so Unione Sovietica 216
City, State ZIP: Turin, Italy 10134
Work Phone: +39 011 3168379
Fax: +39 011 3168877
E-mail: paola.leproni@csi.it
Web: www.csi.it