The "Services To Enterprises" Project

An Experience of Data Quality Improvement in Italian Public Administration

Monica Scannapieco
University of Rome "La Sapienza"
IASI-CNR
monscan@dis.uniroma1.it

Carlo Batini
University of Milan "Bicocca"
Authority for IT in Public Administration
batini@unimib.it
batini@aipa.it

Pietro Aimetti
Gruppo Clas
p.aimetti@gruppoclas.it

Claudio Gagliardi
Unioncamere
claudio.gagliardi@unioncamere.it

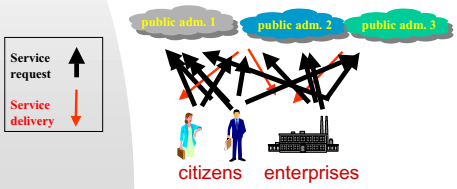
Executive Summary: The paper presents an Italian project in which real problems related to data quality in cooperative information systems have been addressed. The project was realized in the context of the Italian e-Government initiative.

Outline

- Cooperative Information Systems supporting the Italian Public Administration
- Data Quality in Cooperative Information Systems
- The "Services to Enterprise" Project
  - The Data Cleaning Process
  - A Publish&Subscribe (P&S) Architecture
- Final Considerations

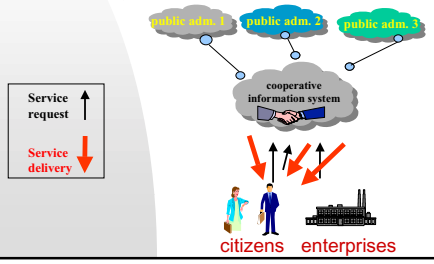
Present Scenario

- Italian Public Administrations do not communicate each other => Unbalanced interactions of Citizens and Enterprises with Public Administrations for service provision

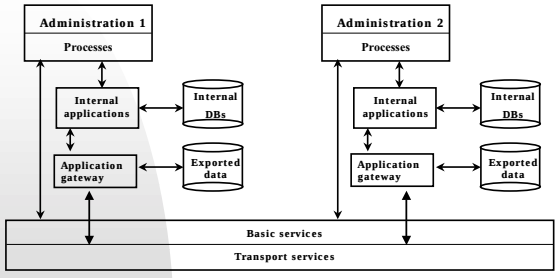


Future Scenario

- Public Administrations communicate through a Cooperative Information System (CIS)
- Balancing of interactions for service provision



The Architecture of the Italian CIS



Data Quality & CIS

- Data quality is an important issue for CIS
- Enables cooperation: if the quality of data is "certified", one organization requests data from another (otherwise not!)
- Data replication can be exploited to improve quality through comparisons of different copies of the same data
- If not addressed, circulation of low quality data may imply a deterioration of the global quality of services

7th International Conference on Information Quality (IQ-2002)

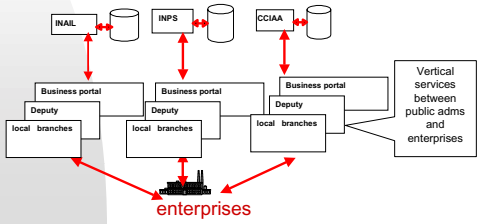
The "Services to Enterprises" Project

- Main Objective: Improving the quality of enterprise-related information by
  - Reconciliation and cleaning of common data that identify and locate enterprises stored in different administrative databases
  - Maintaining the data quality level obtained through the reconciliation and cleaning activity

7th International Conference on Information Quality (IQ-2002)

Starting Scenario

- National insurance agency (INPS)
- National industrial accidents agency (INAIL)
- Chambers of commerce (CCIAA)



7th International Conference on Information Quality (IQ-2002)

Data Sources: 8 different databases

- INPS: 5 databases
- INAIL: 1 databases
- CCIAA : 2 databases

	CCIAA	INPS	INAIL
Number of records	7.098.493	6.110.047	7.065.365
Total number of records		20.273.905	

7th International Conference on Information Quality (IQ-2002)

Main Common Data

- VAT code
- corporate name
- related chamber of commerce
- corporate domicile
- area of business
- production units location
- start of business
- corporate property
- .....

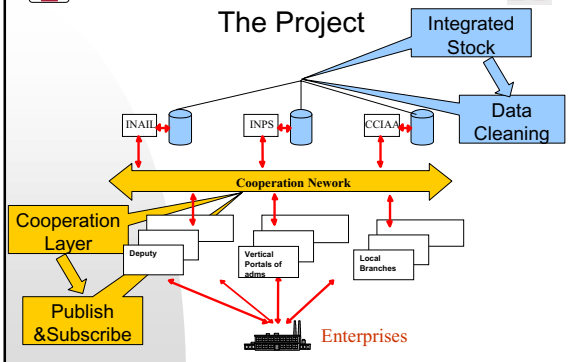
7th International Conference on Information Quality (IQ-2002)

Costs

- High Costs
  - Data management is heavy for enterprises that have to communicate it and for public administrations that have to gather and store it
  - Costs related to creation and modification events have been estimated in ~180 millions of euro for each year
- Low data quality
  - 40% of data not aligned in all the databases
  - Low currency. Average update frequency: 3 months
  - Data quality verification estimated in ~ 25 millions of euro for each year
  - Reduced revenues to be estimated

7th International Conference on Information Quality (IQ-2002)

The Project



7th International Conference on Information Quality (IQ-2002)

The Data Cleaning Process

- Pre-elaboration of data sources: normalization and standardization activities
- Record Linkage : linking of records stored in the different databases and related to the same enterprise (or more generally "economic agent")
- Analysis of not linked records
- Correction of data in each database

7th International Conference on Information Quality (IQ-2002)

Record Linkage

- Different techniques applied in sequence in order to refine the result in the linkage process

Linkage based on previous knowledge

↓

Linkage based on "certain" key:
- intra-archive with exact key matching
- inter-archives with exact key matching
- inter-archives with partial key matching

↓

Probabilistic Linkage

↓

Complex cases analysis

↓

Final Quality Checks

7th International Conference on Information Quality (IQ-2002)

Results of Linkage based on Certain Key

	Links	Non links	% Links
Database 1 CCIAA	5.037.684	1.573.871	76,2
Database 1 INAIL	4.735.485	2.329.880	67,0
Database 1 INPS	1.545.313	30.483	98,1
Database 2 INPS	541.206	39.171	93,3
Database 3 INPS	245.352	137.088	64,2

7th International Conference on Information Quality (IQ-2002)

Results of Probabilistic Linkage

	Links	Non links	% Links
Database 1 CCIAA	5.092.450	1.519.105	77,0
Database 1 INAIL	4.750.115	2.315.250	67,2
Database 1 INPS	1.550.005	25.791	98,4
Database 2 INPS	541.269	39.1108	93,3
Database 3 INPS	247.041	135.399	64,6

7th International Conference on Information Quality (IQ-2002)

Complex Cases Analysis

	Links	Non links	% Links
Database 1 CCIAA	5.077.149	1.534.406	76,8
Database 1 INAIL	4.713.474	2.351.891	66,7
Database 1 INPS	1.545.426	30.370	98,1
Database 2 INPS	535.810	44.567	92,3
Database 3 INPS	246.692	135.748	64,5

7th International Conference on Information Quality (IQ-2002)

Final Results of Record Linkage

	Total	Links	Non links	% Links	% Non links
Database 1 CCIAA	6.611.555	5.077.146	1.539.409	76,7	23,3
Database 1 INAIL	7.065.365	4.705.403	2.359.962	66,6	33,4
Database 1 INPS	1.575.796	1.542.124	33.672	97,9	2,1
Database 2 INPS	580.377	533.234	47.143	91,9	8,1
Database 3 INPS	382.440	246.413	136.027	64,4	35,6

7th International Conference on Information Quality (IQ-2002)

Analysis of Not Linked Records

- Some not linked records derive from that related enterprises have close down

	Total	% Non links	%Non links living enterprises
Database 1 CCIAA	6.611.555	23,3	19,0
Database 1 INAIL	7.065.365	33,4	1,2
Database 1 INPS	1.575.796	2,1	0,2
Database 2 INPS	580.377	8,1	8,0
Database 3 INPS	382.440	35,6	16,9

7th International Conference on Information Quality (IQ-2002)

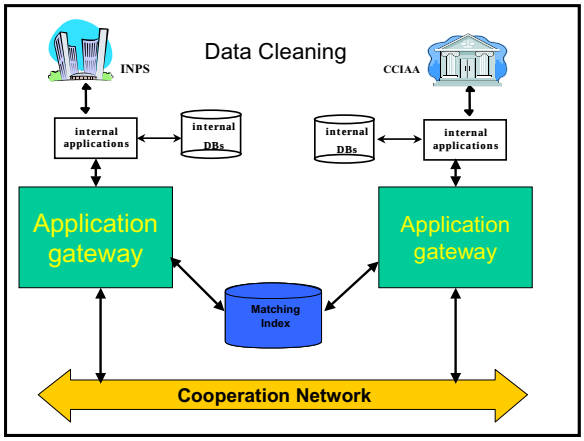
Analysis of Not Linked Records

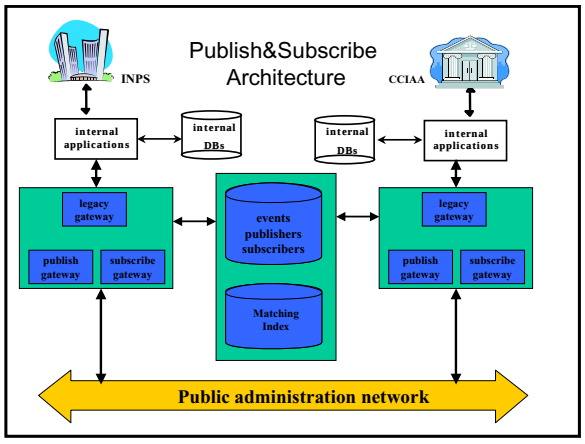
- Limitation of the scope of the project by eliminating records stored in one of the project databases but not interesting for the project
- Codification of domain specific rules in order to make an automatic detection of records to be eliminated

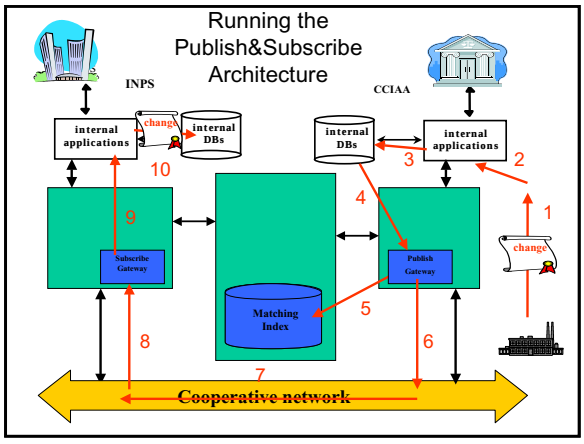
7th International Conference on Information Quality (IQ-2002)

Publish&Subscribe Cooperation Paradigm

- Cooperation necessary in order to maintain Data Quality levels
- A Publish&Subscribe cooperation paradigm allows to express interest for specific events and to be notified when such events occurr
- As a result of the data cleaning phase, a matching index realized in order to maintain the coupling of records among the administrations

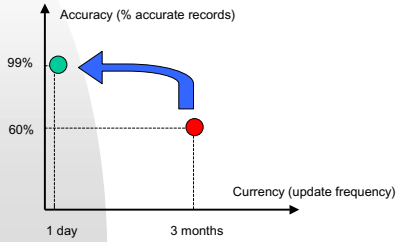






Data Quality Benefits of the Project

- Accuracy: From ~60 % for all the DBs to 99%
- Currency: From an average update frquency of 3 months to 1 day



Final Considerations

	Data Cleaning (to recover the past)	Publish&Subscribe (for maintaining in the future)
Data based Methods	X	
Process based Methods		X

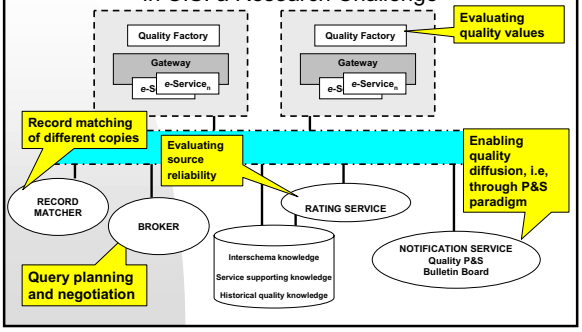
Final Considerations: Data based methods

- Traditional cleaning activities, but...
- Combined application of different techniques in an incremental way
- A phase of linkage based on previous knowledge
- A specific phase for not linked records
- Matching Index to support specific requirements of inter-organizational environments

Final Considerations: Process Based Methods

- Reengineering of inter-organizational cooperative processes “quality-driven”
- The P&S communication implies a mapping of processes into events that administrations exchange each other

A General Architecture Supporting Data Quality in CIS: a Research Challenge



References

- C. Batini and M. Mecella, “Enabling Italian e-Government Through a Cooperative Architecture" IEEE Computer, vol. 34, no. 2, 2001.
- M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, C. Batini: Managing Data Quality in Cooperative Information Systems. In Proceedings of the Tenth International Conference on Cooperative Information Systems (CoopIS 02), Irvine, CA, 2002.
- M. Scannapieco, V. Mirabella, M. Mecella, and C. Batini, “Data Quality in e-Business" in Proceedings of the CAiSE 2002 Workshop on Web Services, e-Business, and the Semantic Web: Foundations,Models, Architecture, Engineering and Applications (WES 2002), Toronto, Ontario, Canada, 2002.