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

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
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

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

Data Sources: 8 different databases
- INPS: 5 databases
- INAIL: 1 databases
- CCIAA: 2 databases

<table>
<thead>
<tr>
<th></th>
<th>CCIAA</th>
<th>INPS</th>
<th>INAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of records</td>
<td>7,886,493</td>
<td>6,110,047</td>
<td>7,885,365</td>
</tr>
<tr>
<td>Total number of records</td>
<td>20,273,865</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Main Common Data
- VAT code
- Corporate name
- Related chamber of commerce
- Corporate domicile
- Area of business
- Production units location
- Start of business
- Corporate property
- Corporate property

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
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

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

Results of Linkage based on Certain Key

<table>
<thead>
<tr>
<th>Database</th>
<th>Links</th>
<th>Non Links</th>
<th>% Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CCIAA</td>
<td>5,037,004</td>
<td>1,534,971</td>
<td>76,2</td>
</tr>
<tr>
<td>1 INAIL</td>
<td>4,735,465</td>
<td>2,329,880</td>
<td>67,0</td>
</tr>
<tr>
<td>1 INPS</td>
<td>1,545,313</td>
<td>30,483</td>
<td>98,1</td>
</tr>
<tr>
<td>2 INPS</td>
<td>541,206</td>
<td>39,171</td>
<td>93,3</td>
</tr>
<tr>
<td>3 INPS</td>
<td>245,352</td>
<td>137,088</td>
<td>64,2</td>
</tr>
</tbody>
</table>

Results of Probabilistic Linkage

<table>
<thead>
<tr>
<th>Database</th>
<th>Links</th>
<th>Non Links</th>
<th>% Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CCIAA</td>
<td>5,092,450</td>
<td>1,519,105</td>
<td>77,0</td>
</tr>
<tr>
<td>1 INAIL</td>
<td>4,750,115</td>
<td>2,315,250</td>
<td>67,2</td>
</tr>
<tr>
<td>1 INPS</td>
<td>1,550,005</td>
<td>25,791</td>
<td>98,4</td>
</tr>
<tr>
<td>2 INPS</td>
<td>541,289</td>
<td>39,110</td>
<td>93,3</td>
</tr>
<tr>
<td>3 INPS</td>
<td>247,041</td>
<td>133,369</td>
<td>64,6</td>
</tr>
</tbody>
</table>

Complex Cases Analysis

<table>
<thead>
<tr>
<th>Database</th>
<th>Links</th>
<th>Non Links</th>
<th>% Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CCIAA</td>
<td>5,077,149</td>
<td>1,534,406</td>
<td>76,8</td>
</tr>
<tr>
<td>1 INAIL</td>
<td>4,713,474</td>
<td>2,351,891</td>
<td>66,7</td>
</tr>
<tr>
<td>1 INPS</td>
<td>1,545,426</td>
<td>30,370</td>
<td>98,1</td>
</tr>
<tr>
<td>2 INPS</td>
<td>535,810</td>
<td>44,567</td>
<td>92,3</td>
</tr>
<tr>
<td>3 INPS</td>
<td>246,692</td>
<td>135,748</td>
<td>64,5</td>
</tr>
</tbody>
</table>

Final Results of Record Linkage

<table>
<thead>
<tr>
<th>Database</th>
<th>Total Links</th>
<th>Total Non Links</th>
<th>% Links</th>
<th>% Non Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CCIAA</td>
<td>6,611,555</td>
<td>5,077,149</td>
<td>78,7</td>
<td>21,3</td>
</tr>
<tr>
<td>1 INAIL</td>
<td>7,085,365</td>
<td>4,750,115</td>
<td>67,2</td>
<td>32,8</td>
</tr>
<tr>
<td>1 INPS</td>
<td>1,550,005</td>
<td>25,791</td>
<td>98,4</td>
<td>1,6</td>
</tr>
<tr>
<td>2 INPS</td>
<td>541,289</td>
<td>39,110</td>
<td>93,3</td>
<td>6,7</td>
</tr>
<tr>
<td>3 INPS</td>
<td>247,041</td>
<td>133,369</td>
<td>64,6</td>
<td>35,4</td>
</tr>
</tbody>
</table>
Analysis of Not Linked Records

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

<table>
<thead>
<tr>
<th>Database</th>
<th>Total</th>
<th>% Non Links</th>
<th>% Other links living enterprises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database 1 CCIAA</td>
<td>6,611,985</td>
<td>23.3</td>
<td>19.0</td>
</tr>
<tr>
<td>Database 1 INAIL</td>
<td>7,065,305</td>
<td>33.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Database 1 INPS</td>
<td>1,575,796</td>
<td>2.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Database 2 INPS</td>
<td>583,377</td>
<td>8.1</td>
<td>8.0</td>
</tr>
<tr>
<td>Database 3 INPS</td>
<td>382,440</td>
<td>35.6</td>
<td>18.9</td>
</tr>
</tbody>
</table>

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

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 occur.
- As a result of the data cleaning phase, a matching index realized in order to maintain the coupling of records among the administrations

Running the Publish&Subscribe Architecture

Cooperative network
Data Quality Benefits of the Project
- Accuracy: From ~60% for all the DBs to 99%
- Currency: From an average update frequency 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:
- 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:
- 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