ACTIVE METADATA TO PRODUCE ADMINISTRATIVE DATA QUALITY DOCUMENTATION

GRAZIA DI BELLA
Istat | Directorate Data Collection

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
Metadata in National Institute of Statistics

Structural, referential and process metadata that traditionally live alongside the data contain a lot of information which, if correctly managed, can also be reused for other purposes.

In this sense, metadata can be made active and “smart”.

In this presentation I will tell you the Istat experience about the acquisition and processing of administrative data. Far-sighted management of the process metadata has made it possible to produce a quality reporting system (QRCA - Quality Report Card of Administrative data) without additional costs.
Agenda

The Istat scenario about Administrative Data Collection
Reusing metadata
Standardizing processes
Reporting quality standards
The Quality Report Card of Administrative data
Some QRCA functions
Conclusions
The Istat scenario about Administrative Data Collection

In Istat, following the modernization process, a Data Collection Directorate was established with the task of centrally acquiring data both from surveys and from administrative sources.

- **Survey data** are managed by multiple systems due to the different types of survey methods or also due to the still ongoing transition towards a complete IT harmonization of collection processes.

- In the **Administrative data** sector, since a centralization process was already active, the collection and treatment process is now almost centralized and standardized. The data that are not included in the central process are mostly macrodata for the exclusive use of a single process or data that requires specific acquisition IT technologies.
Some figures about the use of Admin Data for statistical purposes

Nearly 60% of statistical processes use Administrative data (2023).

They are acquired from 60 Administrative source holders and from 225 sources; in 2023, 712 supplies were acquired (some on an annual basis, others on a quarterly, monthly basis, etc.).

To centrally managed Administrative data, two IT tools have been developed:

→ an acquisition IT system ARCAM
→ and a treatment IT system SIM
Task

This large amount of data that Istat acquires from outside must be necessarily appropriately documented to guarantee the quality, support the administrative data management staff and the statisticians who use them in their production processes or that want to innovate the process and look for available sources.

The information to be reported is extensive and the basic scenario has the following characteristics:

• the quality indicators must be produced for a number ranging for about 700 supplies per year
• the characteristics of administrative datasets have a high variability in terms of format, content, data structure
• administrative datasets are often very large in terms of bytes
• when personal data are present, it is necessary to operate in compliance with the data protection legislation
Centralization opens the door to the possible reuse of process metadata.

The process metadata are very interesting:

**ARCAM System**
- Metadata to manage relationships with holders of Administrative data
- Metadata to manage the acquisition processes

**SIM System**
- Metadata to manage the ETL procedures
- Metadata for the pre-treatment processes in order to comply with the legislation on the protection of personal data while respecting the needs of production (pseudonymization procedures)
Towards the reuse of metadata for the production of the QRCA

With the aim of reusing and making metadata "active“, the two systems - ARCAM and SIM -, were made interoperable by creating primary key transcoding tables that identify the administrative sources and reorganizing some entities and relationships.

The transcoding tables connecting the two systems needs small updates when new datasets or new sources are acquired.

In this way it is possible to follow the life cycle of the administrative data from the statistics needs specification to the collection and the pre-treatment (GSBPM - Generic statistical business process model representation).

The second operation concerned the standardization of metadata which, in the original systems, has the characteristics necessary to carry out the designated functions. For the reuse an overall harmonization was made with respect to the new statistical quality reporting requirements.

Obviously these operations were performed without impacting the primary functions of the metadata.
As regards the new requirements for quality reporting, international statistical standards and specific international references to the definition of the administrative data quality used for statistical purposes were considered. Last but not least, the Istat needs.

For example:


The BLUE-ETS international project

→ 107 theoretical measures have been defined
QRCA strategy

Strategy: Reusing metadata to produce a quality reporting system that automatically updates whenever a new dataset is requested, acquired or processed, or the contents of the acquired datasets change in terms of units or variables.

In this way the QRCA reports are updated in real time by reading the information from the process metadata of the IT data management tools.

Among the 107 measures defined to report on quality of Administrative data, those for which information was already available were implemented:

- currently 60 measures and 19 reports for each Administrative source
QRCA sources

The information that currently feeds the QRCA comes from:

- the ARCAM acquisition system (metadata about relationships with data holders, acquisition processes)

- the SIM microdata integrated system (data description, metadata and macrodata of the treatment processes)

- SIM Oracle data dictionary (columns of the tables and other resources)

- the official statistics planning system - National Statistical Program (association between statistical processes and administrative data used)

- the Istat's output quality reporting system (Sidi-Siqual) (relationships between statistical processes and European legislation)
The QRCA is available on the institute's intranet at qrca.istat.it

- Website built by a generalized Java application
- Visual Analytics platform of BI - MicroStrategy
- An Oracle DB with:
  - Views of the relevant Tables selected by the DBs of the source systems
  - own transcoding tables
  - other tables of processed data that help Microstrategy for the calculation of the measures

Colleagues who want to access the data must forward a specific request following which the specific Views are made available by SIM, in compliance with the legislation on the personal data protection.
## Quality framework

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>METADATA</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic information</td>
<td>Administrative objects</td>
<td>Status of supplies</td>
</tr>
<tr>
<td></td>
<td>Data provided</td>
<td>Temporal aspects</td>
</tr>
<tr>
<td></td>
<td>Views for SIM users</td>
<td>Timeliness and Punctuality</td>
</tr>
<tr>
<td></td>
<td>SIM tables</td>
<td>Record number</td>
</tr>
<tr>
<td></td>
<td>Data structures in SIM</td>
<td>Valued fields</td>
</tr>
<tr>
<td></td>
<td>Classification provided</td>
<td>Frequencies</td>
</tr>
<tr>
<td></td>
<td>Connections between classifications</td>
<td>Missing decodings</td>
</tr>
<tr>
<td>Relevance and statistical uses</td>
<td>Administrative classifications</td>
<td>Integrability/Data integration</td>
</tr>
<tr>
<td>Other available documentation</td>
<td></td>
<td>Linkability of units</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integration Process Monitoring</td>
</tr>
<tr>
<td>SUMMARY REPORTS</td>
<td>Monitoring of acquisitions</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Summary tables of the acquisition processes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Summary tables of the acquisition processes</td>
<td></td>
</tr>
</tbody>
</table>
Search functions available

Search for an administrative variable among those documented in QRCA (6,272 variables): typing a keyword, for example “education" displays all the administrative variables that contain that word in the description and the related sources. From here it is possible to access the source quality reports.
Search for an **administrative source** among the 288 documented in QRCA (using keywords of the name of the source or the name of the holder).

For each source, 19 reports and 60 measures are available.
Basic information about the source

**Vehicle fleet**

**Owner:**

**Presence of personal data (special categories)**

**Pre-Treatment (pseudonymisation)**

**Years available (time series)**

---

**Quality Report Card dei dati Amministrativi**

---

**Archivio**: Parco veicolare

**Ente titolare della Fonte**: ACI - AUTOMOBILE CLUB D'ITALIA

**Presenza di dati personali (Reg. UE 2016/679)**

- SI

**Presenza di dati rientranti in particolari categorie (ex sensibili)**

- NO

**Presenza di dati relativi a condanne puniti e reati (ex giudiziari)**

- NO


**Archivio SIM**

**Tuttramento in SIM**

- Trattamento in SIM con assegnazione del Codice SIM a livello di individuo o di unità economica

**Anni disponibili in SIM**


**Codifica degli indirizzi in RSBL**

- 

**Anni in RSBL**

- 

Referente per l'acquisizione DCRD/DD: Eleonora Ciavolo
Relevance of the source

Employers' monthly contribution declaration

Istat processes using the source
Variables of the source

National registry of university students

Metadata -> Administrative variables -> Data provided

<table>
<thead>
<tr>
<th>Progressive Campo</th>
<th>Campo</th>
<th>Descrizione Campo</th>
<th>Formato</th>
<th>Classificazione</th>
<th>Tabella</th>
<th>Tipo Variable</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PROGRESSIVO_ISCRIZIONE</td>
<td>Codice identificativo interno delle iscrizioni</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>ID - Altra variabile identificativa - Cod. univoco personalizzazione</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ID_STUDENTE</td>
<td>Codice identificativo interno dello studente</td>
<td>N</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>ID - Altra variabile identificativa - Cod. univoco personalizzazione</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CODICE_FISCALE</td>
<td>Codice fiscale</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>ID - Variabile anagrafica - Codice fiscale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SESSO</td>
<td>Sesso</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>ID - Variabile anagrafica - Sesso</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>DATA_NASCITA</td>
<td>Data di nascita</td>
<td>D</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>ID - Variabile anagrafica - Data nascita</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>COD_CONSIGLIO_NASCITA</td>
<td>Codice territoriali di nascita</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>ID - Variabile anagrafica - Luogo nascita</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>COD_PAES_NASCITA</td>
<td>Codice etico di nascita</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>ID - Variabile anagrafica - Luogo nascita</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>DATA_PRIMA_AMM</td>
<td>Data di prima immatricolazione</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>TEM - Variabile relativa all'università</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ANNO_ACC_PRIMA_1MM</td>
<td>Anno accademico di primo immatricolazione (anno iniziale)</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>TEM - Variabile relativa all'università</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CODICE_TIPO_INERIS</td>
<td>Tipo di ingresso in insegnamento</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>TEM - Altra variabile tematica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>ANNO_DIPLOMA</td>
<td>Anno di conseguito del diploma</td>
<td>N</td>
<td>MIUR_UNIVERSITA_ISCRIZIONI</td>
<td>TEM - Variabile relativa all'università</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CODICE_TITOLO_DIPLOMA</td>
<td>Codice Titolo Diploma</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>TEM - Variabile relativa all'università</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>CODICE_CENTRO_DIPLOMA</td>
<td>Codice istituto</td>
<td>A</td>
<td>MIUR_UNIVERSITA_ISCRITTI</td>
<td>TEM - Variabile relativa all'università</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Monitoring of data provided

Pensioners' register

Data -> Monitoring the acquisition of data supplies

Data -> Technical Checks -> Variable completeness

Quality Report Card dei dati Amministrativi

<table>
<thead>
<tr>
<th>Archivio</th>
<th>Casellario dei pensionati e dei trattamenti pensionistici</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ente</td>
<td>INPS - ISTITUTO NAZIONALE DELLA PREVIDENZA SOCIALE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nome Fornitura composto</th>
<th>Progressivo nell'anno</th>
<th>Periodicità</th>
<th>Dati al (gg/mm/aaaa)</th>
<th>Anno di Trattamento</th>
<th>Trattamento previsto</th>
<th>Stato Monitoraggio</th>
<th>Data Stato Monitoraggio</th>
<th>Data Min Ricezione Fornitura</th>
<th>Data Max Ricezione Fornitura</th>
<th>Modalità di Acquisizione</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casellario dei pensionati e dei trattamenti pensionistici</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2023</td>
<td>Trattamento in SIM con assegnazione del Codice SIM a livello di individuo</td>
<td>In fase di richiesta all'ente</td>
<td>01/09/2024</td>
<td>30/09/2024</td>
<td>CANALE SFTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>Trattamento in SIM con assegnazione del Codice SIM a livello di individuo</td>
<td>Disponibile con codici SIM</td>
<td>12/10/2023</td>
<td>01/09/2023</td>
<td>30/09/2023</td>
<td>CANALE SFTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>Trattamento in SIM con assegnazione del Codice SIM a livello di individuo</td>
<td>Disponibile con codici SIM</td>
<td>13/10/2022</td>
<td>01/09/2022</td>
<td>30/09/2022</td>
<td>CANALE SFTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Trattamento in</td>
<td>Disponibile con codici SIM</td>
<td>15/11/2021</td>
<td>01/10/2021</td>
<td>31/10/2021</td>
<td>CANALE SFTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Technical checks of the source

Temporary workers

Data-> Technical Checks -> Number of records

Data-> Technical Checks -> Variable completeness
Integrability of the source

National registry of school students

Data -> Integrability

Linkage variable available and quality

Monitoring the integration process (pseudonymisation)
Conclusions

Istat has experienced the advantages of reusing process metadata for the production of quality reports.

The idea is to extend the use of this strategy as much as possible.

In general, in a centralize it is important to expand the connections between metadata to improve interoperability between datasets even coming from different types of sources.

From this perspective, a multi-source input quality documentation system is being designed.
thanks

GRAZIA DI BELLA | grazia.dibella@istat.it