Paris ,11-12 April 2024 France.
Conference on Smart Metadata for Official Statistics, 2024 edition
Session 6

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

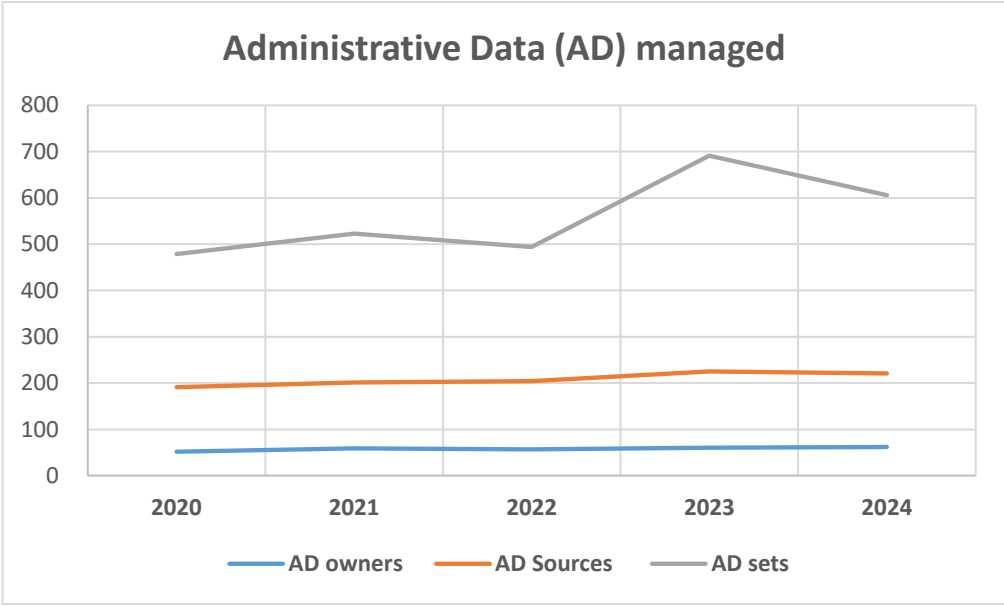
Nearly 60% of statistical processes use Administrative data (2023).

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

Process metadata for Admin data management

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.

Administrative quality reporting

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 Single Integrated Metadata Structure (SIMS 2.0) SIMS-2-0-Revised-standards-November-2015-ESSC-final.pdf (europa.eu)

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

QRCA component

Quality Report Card of Administrative data

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

---

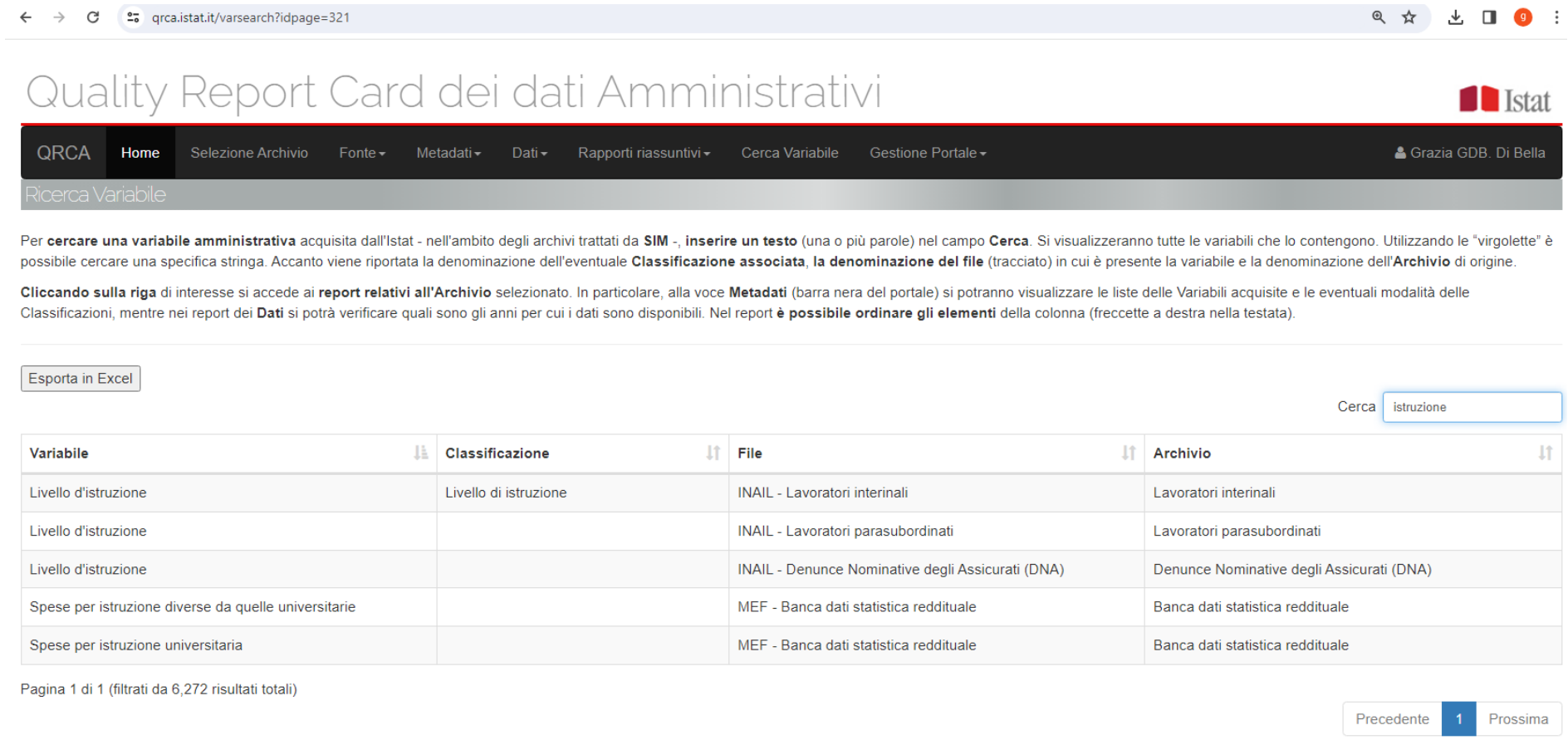
Colleagues who want to access the data must forward a specific request following which the specific Views are made available by SIM, in compliance with the legislation on the personal data protection.

Quality framework

SOURCE	METADATA			DATA	
Basic information	Clarity and interpretability of data	Administrative objects		Status of supplies	
		Administrative variables	Data provided	Temporal aspects	Timeliness and Punctuality
			Views for SIM users	Technical checks - Compliance	Record number
			SIM tables		Valued fields
			Data structures in SIM		Frequencies
		Administrative classifications	Classifications provided		Missing decodings
Relevance and statistical uses			Connections between classifications	Integrability/Data integration	Linkability of units
	Other available documentation				Integration Process Monitoring
				Completeness	
SUMMARY REPORTS	Monitoring of acquisitions Summary tables of the acquisition processes Summary tables of the acquisition processes				

Search functions available

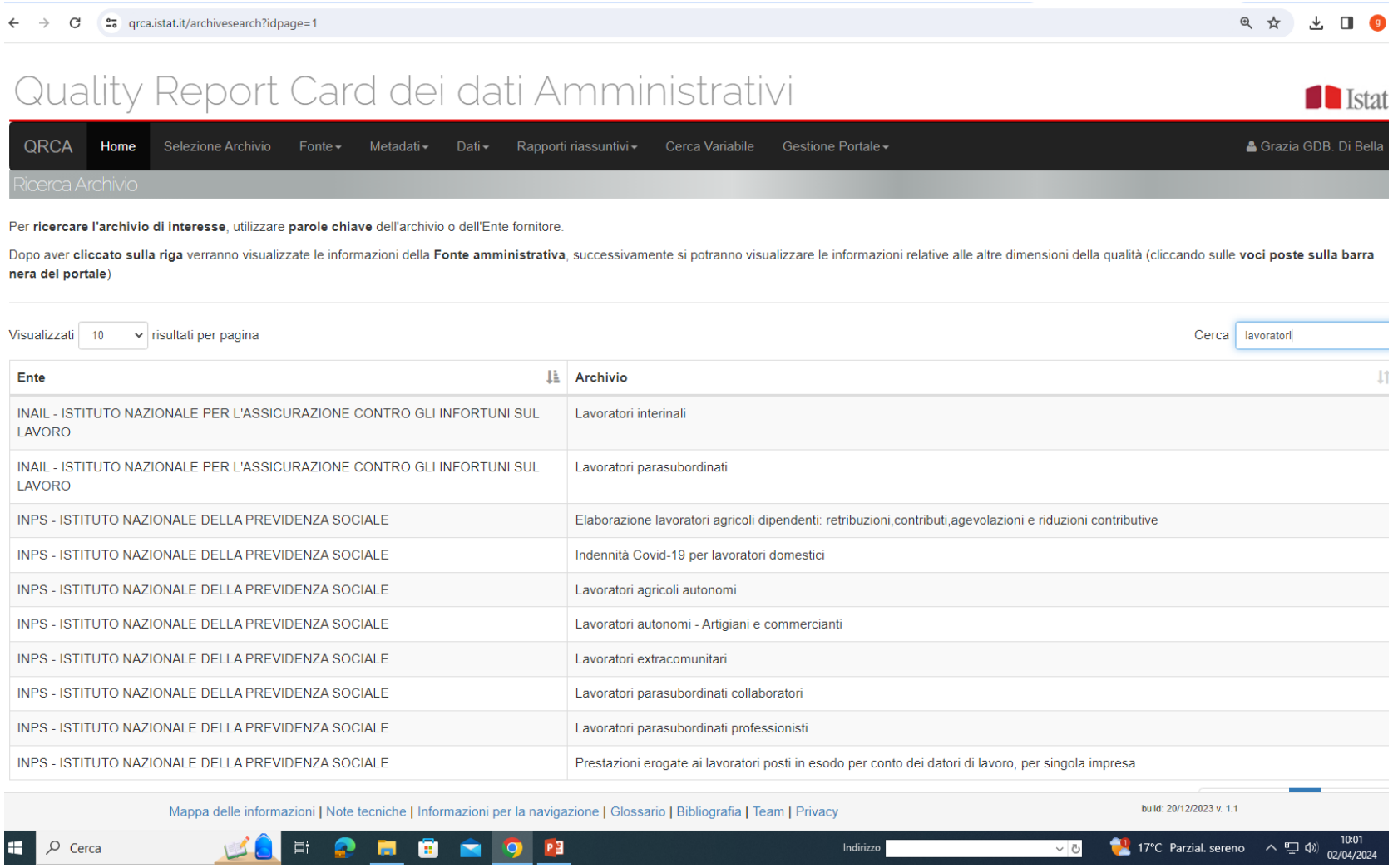
Search for an **administrative variable** among those documented in QRCA (6,272 variables): typing a keyword, for example “education" displays all the administrative variables that contain that word in the description and the related sources. From here it is possible to access the source quality reports



Search functions available

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



Istat

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

QRCA | Home | Selezione Archivio | Fonte | Metadati | Dati | Rapporti riassuntivi | Cerca Variabile | Gestione Portale

Fonte -> Informazioni di base

Si riportano alcune informazioni relative alle **caratteristiche** della Fonte amministrativa e dell'Archivio da essa derivato.
La **disponibilità** degli archivi è riferita al completamento del processo di trattamento previsto per l'Archivio; per Archivi con più forniture in un anno, indica che almeno una fornitura dell'anno è disponibile (per avere informazioni sul monitoraggio dello stato delle forniture, consultare il report Dati -> Stato delle Forniture). La serie storica degli anni disponibili può essere troncata a sinistra qualora l'archivio sia stato acquisito prima dell'anno 2017 e non sia stato trattato in SIM. I **trattamenti** documentati sono: • nel Sistema Integrato dei Microdati (SIM): gestione dei metadati, assegnazione dei codici agli Individui e alle Unità economiche (pseudonimizzazione secondo il Reg. UE 2016/679); • nel Registro Statistico di Base dei Luoghi (RSBL): assegnazione del CUI, georeferenziazione e geocodifica degli indirizzi. Per informazioni più dettagliate sui trattamenti cfr. le Note tecniche nei footer.
Dettagli in merito all'acquisizione delle forniture della Fonte possono essere richieste al **referente** indicato. Un campo Note riporta informazioni a livello di Archivio.

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 penali e reati (ex giudiziari)	NO
Anni Disponibili	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022

Archivio SIM	Trattamento in SIM	Anni disponibili in SIM	Codifica degli indirizzi in RSBL	Anni in RSBL
ACI - Parco veicolare	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022	-	

Referente per l'acquisizione DCRD/RDD: Eleonora Ciocca

Note	

Relevance of the source

Employers' monthly contribution declaration

Istat processes using the source

Quality Report Card dei dati Amministrativi

QRCA | Home | Selezione Archivio | Fonte | Metadati | Dati | Rapporti riassuntivi | Cerca Variabile | Gestione Portale

Fonte -> Rilevanza e usi statistici

L'**indicatore di Rilevanza** fornisce il grado di importanza e pertinenza dell'archivio per la produzione delle statistiche ufficiali. Le misure presentate sono espresse in termini di estensione
Statistico Nazionale) dell'Istat che utilizzano l'archivio e della normativa UE il cui adempimento dipende dall'uso dell'archivio. Le informazioni derivano rispettivamente dal Sistema Informativo de
delle Indagini).
Per completare le informazioni sull'uso statistico dell'archivio per i vari lavori PSN, consultare Il Sistema Informativo sulla Qualità (SIQual) http://siqual-intra.istat.it selezionando il processo produt

Archivio	UNIEMENS - Parte contributiva
Ente	INPS - ISTITUTO NAZIONALE DELLA PREVIDENZA SOCIALE

Numero di lavori Istat del PSN per i quali si dichiara di utilizzare l'archivio per annualità (Fonte: PSN)

Triennio	Annualità	Stato	Data di pubblicazione in Gazzetta Ufficiale	Numero di lavori	Numero di normative UE
2023-2025	2023	In corso di formalizzazione		26	30
2020-2022	2022	In vigore	29/08/2023	26	30
2020-2022	2021	Non più in vigore	21/02/2023	23	28
2020-2022	2020	Non più in vigore	26/05/2022	22	28
2017-2019	2019	Non più in vigore	21/02/2021	18	24
2017-2019	2018	Non più in vigore	16/07/2019	13	20
2017-2019	2017	Non più in vigore	20/03/2018	6	14

Lavori Istat del Psn per i quali si dichiara di utilizzare l'archivio e Normativa UE il cui adempimento dipende dall'uso dell'archivio (Fonti: PSN, SIDI)

Annualità del Psn 2023

Variables of the source

National registry of university students

Metadata -< Administrative variables -> Data provided

Quality Report Card dei dati Amministrativi

QRCA | Home | Selezione Archivio | Fonte | Metadati | Dati | Rapporti riassuntivi | Cerca Variabile | Gestione Portale | Grazia GDB. DI

Metadati -> Chiarezza/Interpretabilità -> Variabili amministrative -> Dati forniti dall'Ente

Sono riportate le **variabili amministrative** richieste all'Ente, esse vengono scelte in base ad un'istruttoria preliminare alla fase di acquisizione.
La struttura dei dati rispecchia le caratteristiche della fonte e può comprendere uno o più file, a cui corrispondono uno o più tracciati. Un filtro permette di selezionare i file con **tracciati attivi**, cioè utilizzati per l'invio corrente dei dati e i file con **tracciati chiusi** (quando presenti), ovvero nella forma precedente a quella attuale. Al tracciato è associato il periodo di validità.
Per ciascun tracciato sono riportate le descrizioni delle variabili, inserite in fase di trattamento, e altre informazioni di interesse. Se presente, cliccare sulla Classificazione amministrativa per visualizzarne le modalità, per uscire dal popup cliccare al di fuori.
Il campo **Tipo variabile** riporta la classificazione delle variabili amministrative **effettuata in SIM** per la gestione delle misure di trattamento dei **dati personali**.

Archivio	Anagrafe nazionale degli studenti universitari - Iscritti e iscrizioni
Ente	MUR - MINISTERO UNIVERSITÀ E RICERCA
Archivi connessi	

Tipo di tracciato: Tracciati attivi

Progressivo Tracciato: 474: MUR - Anagrafe degli iscritti all'universita' (dal 2019)

Progressivo Campo	Campo	Descrizione Campo	Formato	Classificazione	Tabella	Tipo Variabile	Note
1	PROGRESSIVO_ISCRIZIONE	Codice identificativo interno della immatricolazione	A		MIUR_UNIVERSITA_ISCRIZIONI	ID - Altra variabile identificativa - Cod. univoco persona/azienda	
2	ID_STUDENTE	Codice identificativo interno dello studente	N		MIUR_UNIVERSITA_ISCRIZIONI	ID - Altra variabile identificativa - Cod. univoco persona/azienda	
3	CODICE_FISCALE	Codice fiscale	A		MIUR_UNIVERSITA_ISCRITTI	ID - Variabile anagrafica - Codice fiscale	
4	SESSO	Sesso	A		MIUR_UNIVERSITA_ISCRITTI	ID - Variabile anagrafica - Sesso	
5	DATA_NASCITA	Data di nascita	D		MIUR_UNIVERSITA_ISCRITTI	ID - Variabile anagrafica - Data nascita	
6	COD_ISTAT_COM_NASCITA	Codici territoriali di nascita	A			ID - Variabile anagrafica - Luogo nascita	
7	COD_PAESE_NASCITA	Codice stato di nascita	A		MIUR_UNIVERSITA_ISCRITTI	ID - Variabile anagrafica - Luogo nascita	Classificazione Istat Paesi esteri pubblicata sul sito www.istat.it
8	DATA_PRIMA_IMM	Data di prima immatricolazione	A		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Variabile relativa all'istruzione	
9	ANNO_ACC_PRIMA_IMM	Anno accademico di prima immatricolazione (limite inferiore)	A		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Variabile relativa all'istruzione	
10	CODICE_TIPO_INGRESSO	Tipo di ingresso in anagrafe	A		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Altra variabile tematica	
11	ANNO_DIPLOMA	Anno di conseguimento del diploma	N		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Variabile relativa all'istruzione	
12	CODICE_TITOLO_DIPLOMA	Codice Titolo Diploma	A		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Variabile relativa all'istruzione	
13	CODICE_ISTITUTO_DIPLOMA	Codice Istituto Conseguimento Diploma	A		MIUR_UNIVERSITA_ISCRIZIONI	TEM - Variabile relativa all'istruzione	

Monitoring of data provided

Pensioners' register

Data-> Monitoring the acquisition of data supplies

Data-> Technical Checks -> Variable completeness

Quality Report Card dei dati Amministrativi

QRCA | Home | Selezione Archivio | Fonte | Metadati | Dati | Rapporti riassuntivi | Cerca Variabile | Gestione Portale

Dati -> Stato delle Forniture

Archivio	**Casellario dei pensionati e dei trattamenti pensionistici**
Ente	**INPS - ISTITUTO NAZIONALE DELLA PREVIDENZA SOCIALE**

Nome Fornitura composto	Progressivo nell'anno	Periodicità	Dati al (gg/mm)	Anno di Riferimento	Trattamento previsto	Stato Monitoraggio	Data Stato Monitoraggio	Data Min Ricezione Fornitura	Data Max Ricezione Fornitura	Modalità di Acquisizione	Note
Casellario dei pensionati e dei trattamenti pensionistici		Annuale		2023	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo	In fase di richiesta all'ente		01/09/2024	30/09/2024	CANALE SFTP	
				2022	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo	Disponibile con codici SIM	12/10/2023	01/09/2023	30/09/2023	CANALE SFTP	
				2021	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo	Disponibile con codici SIM	13/10/2022	01/09/2022	30/09/2022	CANALE SFTP	
				2020	Trattamento in	Disponibile con codici SIM	15/11/2021	01/10/2021	31/10/2021	CANALE	

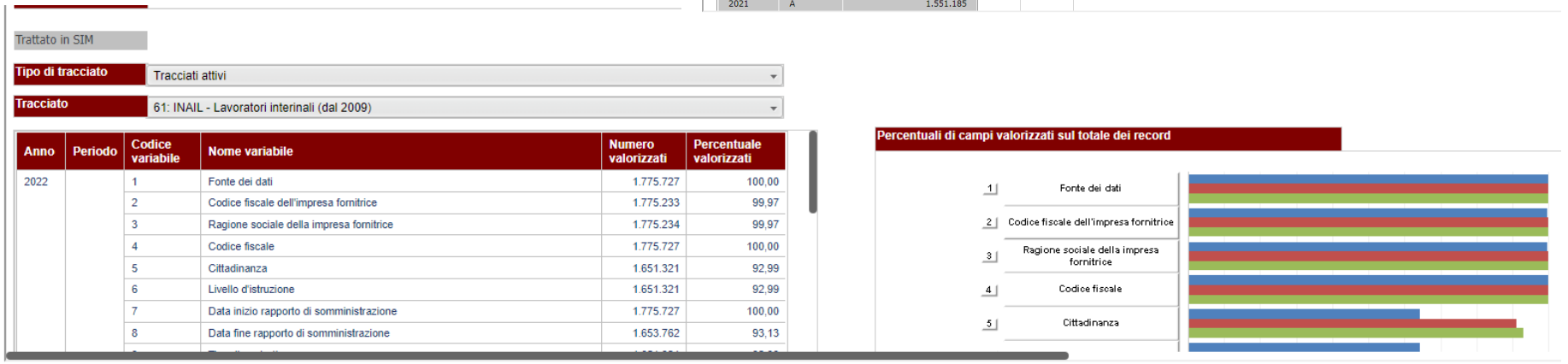
Mappa delle informazioni | Note tecniche | Informazioni per la navigazione | Glossario | Bibliografia | Team | Privacy

Technical checks of the source

Temporary workers

Data-> Technical Checks -> Number of records

Data-> Technical Checks -> Variable completeness



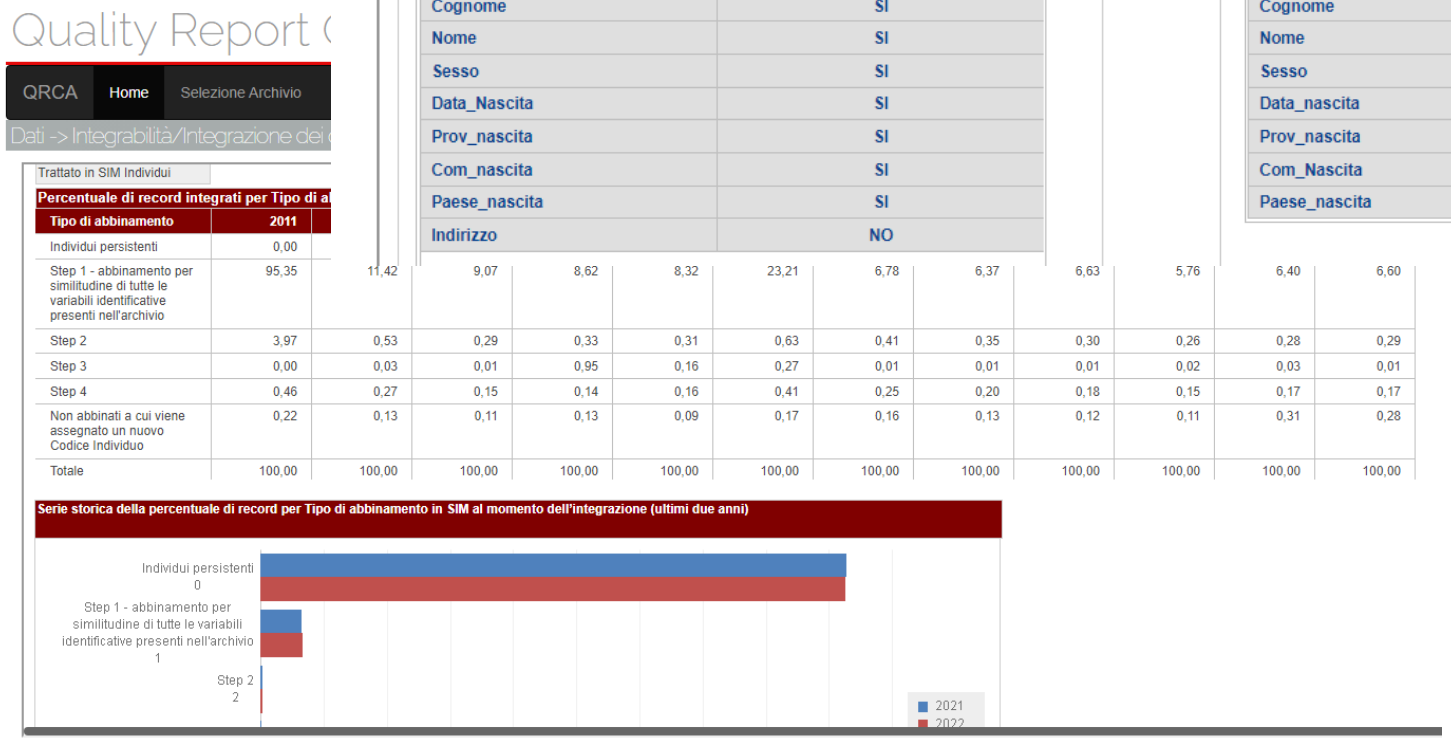
Integrability of the sourc

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

Istat | Istituto Nazionale di Statistica