Istat
Istituto Nazionale di Statistica

IL SISTEMA DI DOCUMENTAZIONE DEI DATI AMMINISTRATIVI IN ISTAT

IL SISTEMA DI DOCUMENTAZIONE DEI DATI AMMINISTRATIVI IN ISTAT

EDIZIONE 2021

Contenuti a cura di: Grazia Di Bella.

Attività editoriali: Nadia Mignolli (coordinamento), Marzia Albanesi, Patrizia Balzano e Alessandro Franzò.
Responsabile per la grafica: Sofia Barletta.

ISBN 978-88-458-2063-2

INDICE

INTRODUZIONE[1]

In questi ultimi anni l'Istat ha profondamente innovato i processi di produzione con lo scopo di ampliare l'offerta informativa e aumentare l'efficienza e la qualità del sistema. In particolare, il progressivo utilizzo di microdati da fonte amministrativa e la costruzione dei Registri statistici (Garofalo, 2016), che permettono una descrizione più dettagliata dei fenomeni in una prospettiva multi-dominio, hanno reso necessario lo sviluppo di nuove definizioni, metodologie e di nuovi strumenti. Per dare una misura di questa trasformazione, basti ricordare che nel Programma statistico nazionale del triennio 2017-2019, aggiornamento 2019, il 41,5 per cento dei lavori statistici dell'Istat dichiarano l'utilizzo di dati da fonte amministrativa. Una percentuale piuttosto elevata considerando anche tale misura come grado di dipendenza della statistica ufficiale da dati prodotti esternamente all'Istituto. Lo scenario presenta, quindi, una cresciuta collaborazione con gli Enti titolari di dati amministrativi, per la maggior parte afferenti alla Pubblica Amministrazione, ed evidenzia la necessità di mantenere una coesione di intenti che garantisca la continuità e la qualità della produzione anche nel rispetto dei Principi del Codice delle statistiche europee[2] (Eurostat, 2017).

Le innovazioni, inizialmente sperimentali, si avviano verso una progressiva fase di standardizzazione sia dal punto di vista gestionale che metodologico. La necessaria documentazione della qualità nei suoi molteplici aspetti, quindi, si va consolidando.

Questa pubblicazione presenta il portale di documentazione dei dati amministrativi acquisiti dall'Istat dalle fonti esterne ed utilizzati nella produzione delle statistiche. L'obiettivo è di documentare la qualità dell'input dei processi che utilizzano dati amministrativi. Il sistema è denominato QRCA - Quality Report Card dei dati Amministrativi e comprende una descrizione del ciclo di vita dei dati amministrativi in Istat e gli indicatori di qualità del processo di acquisizione e pretrattamento statistico dei dati stessi. L'idea della QRCA nasce nell'ambito del progetto internazionale BLUE ETS - Enterprise and Trade Statistics (Daas et al., 2011b, Daas et al., 2011c) e poi si sviluppa in Istat (Cerroni et al., 2014), in linea con gli standard internazionali, (Eurostat, 2017) con lo scopo di adattarsi alle esigenze specifiche dell'Istat e al processo di modernizzazione avviatosi nel 2016.

Di particolare interesse è la strategia di implementazione adottata per la produzione della QRCA che prevede l'utilizzo dei metadati di processo degli IT Tool di gestione dei dati amministrativi e i paradati disponibili, alimentandosi in modo automatico e garantendo un aggiornamento in tempo reale. Il portale della QRCA, accessibile solo agli utenti interni dell'Istat, si basa su un'applicazione Java e utilizza il BI Microstrategy.

1 Questa pubblicazione è stata curata da Grazia Di Bella. Autori: Grazia Di Bella (Premessa, Capitolo 1, Capitolo 2, Paragrafo 3.1, Conclusioni e sviluppi futuri, Glossario, Appendice), Simona Spirito (Paragrafi 3.2, 3.4, 3.5, 3.6.), Grazia Petraccone e Monica Porcelli (Paragrafo 3.3), Raffaella Rosati ha effettuato le elaborazioni sui dati.

2 Diversi principi del Codice delle statistiche europee richiamano aspetti connessi all'uso statistico dei dati amministrativi. I Principi 2 e 5 relativi all'accesso ai dati e alla loro protezione, aspetti strettamente connessi che devono essere affrontati dalle Autorità statistiche, il Principio 8, in particolare gli indicatori 8.6 e 8.7 riguardanti le procedure di collaborazione con gli Enti titolari dei dati amministrativi, il Principio 9 sul *Response burden* che richiama esplicitamente all'uso delle Fonti amministrative(indicatore 9.4), rafforzato dal Principio10 sui Rapporti costi/efficacia per la limitazione al ricorso a indagini dirette (10.3).

Ad oggi sono state implementate le informazioni e gli indicatori i cui dati di base sono disponibili nei sistemi di gestione, nel futuro è previsto di sfruttare ulteriormente le informazioni disponibili e di alimentare i report della QRCA con ulteriori sistemi. Gestire la qualità in Istituto è più facile dal momento in cui i processi si vanno via via standardizzando e l'adozione di specifici IT Tool per la gestione dei dati amministrativi, per la maggior parte dati personali, rende possibile documentare i processi e perfezionare sempre di più il sistema di garanzia della riservatezza, dell'integrità e della qualità dei dati, nel rispetto della normativa sulla protezione dei dati - *by design* e *by default* in un'ottica sistemica.

Nel capitolo 1 si introducono i concetti di base relativi ai dati amministrativi: la terminologia adottata (§1.1) e la loro contestualizzazione in Istat attraverso la descrizione del processo di gestione dei dati amministrativi con lo scopo di organizzare la narrazione della loro documentazione (§1.2). La qualità e il framework teorico di riferimento sono riportati nel paragrafo 1.3. Chiude il primo capitolo la presentazione delle tipologie di utenti della QRCA.

Nel secondo capitolo della pubblicazione si presenta il portale QRCA, gli elementi di carattere metodologico vengono inquadrati e sono descritti gli indicatori di valutazione della qualità.

Nel capitolo terzo viene presentata la strategia adottata, le procedure di manutenzione e controllo periodici del sistema.

Un ultimo paragrafo è dedicato alle conclusioni e agli sviluppi futuri nell'attuale scenario in cui la normativa sulla protezione dei dati richiede di adottare e documentare specifiche misure di garanzia e l'uso dei microdati amministrativi diventa sempre più esteso per la produzione dei Registri statistici dell'Istat.

Un'Appendice al Volume riporta l'insieme complessivo delle misure della qualità dei dati amministrativi definite inizialmente in forma teorica e adattate, in fase di implementazione, alle esigenze informative dell'Istituto; è indicato, inoltre, il loro stato attuale di implementazione nel portale QRCA 2.0. In relazione agli sviluppi futuri indicati, l'insieme delle misure è in progressiva evoluzione.

Il Glossario della QRCA, che elenca i termini legati alla gestione e all'uso dei dati amministrativi per la produzione delle statistiche in Istat, conclude la pubblicazione.

1. LA QUALITÀ DEI DATI AMMINISTRATIVI E GLI OBIETTIVI DELLA QRCA

1.1 La qualità dell'input dei processi di produzione statistica

Con il termine qualità dei dati amministrativi ci si riferisce alla qualità dei dati acquisiti dalle Fonti amministrative ed utilizzati come input dei processi di produzione delle statistiche ufficiali prodotte dall'Istat. Occorre premettere che ci si riferisce alla "qualità statistica" dei dati amministrativi ovvero alle caratteristiche necessarie che i dati di input devono avere per la produzione di un output statistico di qualità, secondo i principi europei (Eurostat, 2017), a seguito di un processo di trattamento o di stima. La qualità statistica del dato amministrativo può essere diversa dalla qualità del dato amministrativo in sé: un dato amministrativo può essere di ottima qualità per le finalità per cui è stato prodotto ma avere un'usabilità statistica più o meno elevata, ad esempio dati che assolvono alle finalità amministrative ma la cui popolazione target amministrativa non coincide con la popolazione target statistica, adozione di concetti e classificazioni dei dati diversi da quelli statistici per cui è necessario operare delle trasformazioni e così via.

Il monitoraggio della qualità dei dati utilizzati come input dei processi è anche un concetto fondamentale perché i dati amministrativi dipendono da adempimenti amministrativi o finalità gestionali che possono mutare nel tempo: mantenere un osservatorio sull'andamento della qualità delle singole forniture acquisite è indispensabile per garantire la produzione statistica e mantenere un elevato standard di qualità del prodotto statistico laddove la dipendenza da queste fonti di dati è sempre crescente.

Il concetto di qualità dei dati di input si inserisce nell'utile approccio di Zhang, 2012 e Reid *et al.*, 2017, in cui la formalizzazione del ciclo di vita dei dati viene esteso ai processi di produzione che utilizzano più fonti e, in particolare, fonti amministrative. In tale visione, l'approccio della QRCA è coerente con la valutazione dell'input delle singole fonti effettuato preliminarmente per ciascuna fonte nell'ambito della fase 1 e in parte della fase 2 relativa all'integrazione tra i dati amministrativi, come si vedrà di seguito.

1.1.1 Gli oggetti della QRCA

Questo paragrafo è necessario per inquadrare esattamente gli oggetti documentati. A questo scopo si fa riferimento al Glossario della QRCA le cui definizioni derivano in gran parte il Glossario dell'ESSnet Admin Data prodotto in seguito ad una ricognizione dei principali glossari internazionali[1].

1 *Glossari internazionali:*
- *Eurostat Concepts and Definitions Database (CODED)* http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODED2&StrLanguageCode=EN
- *OECD Glossary of Statistical Terms* http://stats.oecd.org/glossary/
- *UNdata Glossary (UN Statistics Division)* http://data.un.org/Glossary.aspx
- *Metadata Common Vocabulary SDMX* https://sdmx.org/?sdmx_news=sdmx-glossary

Dal Glossario AdminData sono state selezionate le voci rilevanti per la QRCA, è stata effettuata la traduzione in italiano e sono state inserite delle modifiche per rendere le definizioni più pertinenti all'ambito nazionale. Infine, sono state inserite nuove voci: Archivio amministrativo, Fornitura di dati amministrativi, Dati prodotti a scopo commerciale, Dati prodotti per finalità gestionali. In fondo al testo è riportato il Glossario integrale della QRCA.

Il primo concetto gerarchico da definire è la Fonte amministrativa:

Fonte amministrativa

"Un insieme di dati raccolti e mantenuti per l'attuazione di uno o più regolamenti amministrativi.

In un senso più ampio, qualsiasi fonte di dati che contiene informazioni che non sono raccolte principalmente per scopi statistici".

Dal concetto di Fonte deriva il concetto di Archivio Amministrativo che costituisce il nucleo dell'informazione della QRCA. Esso è definito nel seguente modo:

Archivio amministrativo

"Struttura del dataset amministrativo definita come sottoinsieme di una Fonte amministrativa".

Come già detto, il termine Archivio amministrativo non trova un corrispondente nei Glossari internazionali e, in generale, nella terminologia anglosassone ed è stato inserito perché derivante dalla consuetudine in Istat tra i ricercatori di utilizzare questo termine per indicare l'insieme di dati amministrativi che si utilizzano nel processo di produzione. Quando possibile, l'accordo con il titolare dei dati amministrativi prevede la definizione del tracciato record dei dati ovvero l'elenco delle variabili e degli oggetti amministrativi (eventi o unità) da estrarre dalla Fonte amministrativa presso l'Ente e da inviare all'Istat. La struttura del dataset derivante dall'incontro tra i fabbisogni statistici e i contenuti informativi della Fonte costituisce l'Archivio amministrativo. Un ulteriore elemento che caratterizza questa definizione è il rispetto della normativa per il trattamento dei dati personali: nella definizione di Archivio si applica il principio di minimizzazione dei dati che sono "adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per le quali sono trattati" (Capo II, art. 5 del Regolamento UE, 2016/679) laddove l'acquisizione dei dati da parte dell'Istat riguarda solo il sottoinsieme di interesse.

Il terzo oggetto fondamentale è la *Fornitura dei dati amministrativi* intesa come:

"Uno o più dataset amministrativi ricevuti dall'Istituto nazionale di statistica da parte del Fornitore di dati amministrativi in seguito ad una specifica richiesta".

Dove il Dataset amministrativo è così definito:

"Insieme strutturato di dati estratti da una Fonte amministrativa, prima di qualsiasi trattamento o validazione da parte dell'Istituto nazionale di statistica".

Si tratta, quindi, dei dati effettivi che vengono regolarmente acquisiti dall'Istat.

I data set della fornitura sono caratterizzati da un preciso riferimento temporale, inoltre ad essi può anche essere associata una data di estrazione dalla Fonte amministrativa o più generalmente una data di creazione. Nel caso in cui una stessa fornitura venga prodotta in tempi diversi, si può parlare di "Fornitura di dati provvisori" o "Fornitura di dati definitivi" o di "Aggiornamento dei dati della fornitura", ovviamente sempre con la specificazione della data di creazione. Un attributo della fornitura è la data di consegna".

La definizione di Dati amministrativi completa l'elenco degli oggetti principali descritti nella QRCA.

"Dati derivati da una Fonte amministrativa, prima di ogni processo di validazione da parte dell'Istituto nazionale di statistica" e in senso più ampio, "Dati non raccolti principalmente per scopi statistici".

Quindi, un Archivio amministrativo è un sottoinsieme della Fonte amministrativa e la Fornitura di dati amministrativi è l'insieme dei dati definiti dalla struttura dell'Archivio che viene acquisito periodicamente.

La QRCA documenta queste entità cercando di fornirne le caratteristiche principali in termini descrittivi e qualitativi.

1.2 Il ciclo di vita dei dati amministrativi in Istat

L'attività di acquisizione e di pretrattamento dei dati amministrativi in Istat è centralizzata in un'unica struttura con lo scopo di supportare i processi di produzione ed evitare duplicazioni di attività.

Tale centralizzazione, avviatasi nel 2007, ha permesso nel tempo di ridurre il sempre più crescente carico informativo presso i titolari dei dati amministrativi, grazie al coordinamento della raccolta dei fabbisogni dei vari processi statistici.

In precedenza, ciascun processo di produzione aveva dei propri referenti presso gli Enti fornitori e adottava propri canali di acquisizione dei dati. Questa impostazione, con l'aumentare delle forniture, è diventata insostenibile per problemi di sicurezza, mancanza di condivisione delle attività di acquisizione ed eccessivo carico sui fornitori, in special modo di alcuni fornitori chiave.

La centralizzazione delle attività di acquisizione ha, inoltre, permesso di avviare una fase di innovazione degli strumenti tecnologici di acquisizione mettendo in sicurezza i dati e garantendo una standardizzazione dei processi. Infine, è stato ottimizzato il Programma annuale di acquisizione delle forniture di dati amministrativi sia rispetto ai contenuti che rispetto alle tempistiche di acquisizione. Per questo ultimo aspetto, occorre considerare che la tempestività delle statistiche vincola l'usabilità dei dati amministrativi e, ad esempio, se le statistiche congiunturali hanno bisogno di una tempestività di acquisizione maggiore infra-annuale, anche a discapito della completezza dei dati, le statistiche strutturali possono programmare acquisizioni a cadenza annuale, ne consegue che il coordinamento dei fabbisogni deve tener conto delle esigenze e delle disponibilità anche in relazione alla qualità dei dati, nell'ottica di una ottimizzazione complessiva del sistema.

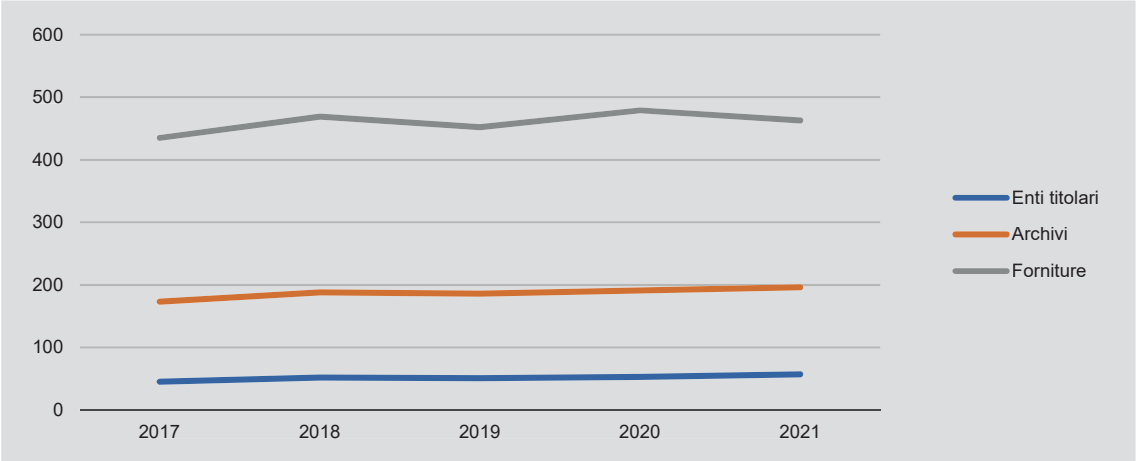
Dopo il periodo di forte crescita avviatosi circa 20 anni fa e sollecitato dalla fase esplorativa delle nuove fonti e la sempre più proficua collaborazione degli Enti titolari nell'obiettivo condiviso di fornire informazione di qualità al Paese, negli ultimi anni si assiste ad una stabilizzazione del numero di Enti con un leggero ma progressivo aumento del numero degli Archivi e delle Forniture come mostrato in Tavola 1.1 e in Figura 1.1.

Tavola 1.1 - Enti titolari, Archivi, Forniture inviate annualmente. Anni 2017-2021

ANNO	Enti titolari	Archivi	Forniture
2017	45	173	435
2018	52	188	469
2019	51	186	452
2020	53	191	479
2021	57	196	463

Fonte: Istat, elaborazione su dati QRCA

Figura 1.1 - Evoluzione del numero di Enti titolari, Archivi, Forniture. Anni 2017-2021



Fonte: Istat, QRCA

D'altra parte, nel Programma statistico nazionale (Psn) cresce la quota percentuale dei lavori statistici programmati dall'Istat che dichiarano l'utilizzo di dati amministrativi. Nella Tavola 1.2 si riportano i valori relativi al triennio di programmazione 2017-2019. Dal 2017 al 2019, in soli due anni, l'incidenza è quasi raddoppiata e nel 2019 più del 40 per cento dei lavori statistici dichiara l'utilizzo di dati amministrativi.

Tavola 1.2 - Quota in percentuale dei lavori Istat del Psn che utilizzano dati amministrativi. Anni 2017-2019

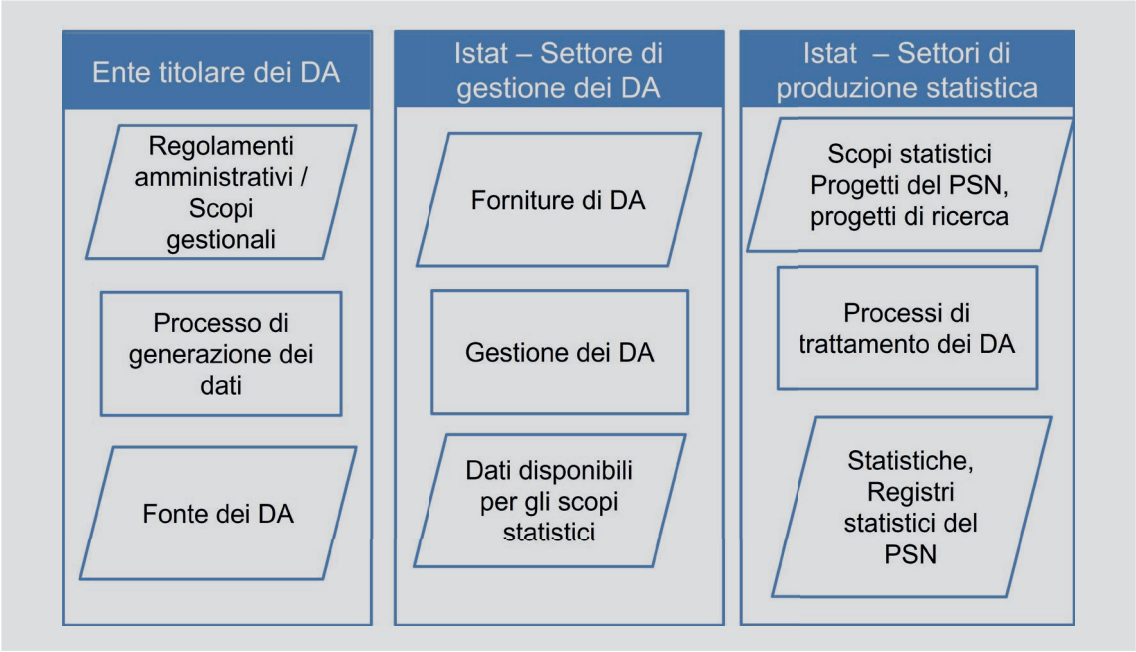
ANNO	Lavori Istat programmati nel PSN	Lavori Istat del PSN che utilizzano dati amministrativi	Incidenza
2017	325	70	21,5
2018	319	102	32,0
2019	318	132	41,5

Fonte: Sistan

Per inquadrare i processi di gestione e trattamento dei dati amministrativi, è utile allargare la prospettiva comprendendo l'intero ciclo di vita dei dati amministrativi, schematizzato in Figura 1.2. Nella prima colonna sono riportate le attività che avvengono presso l'Ente titolare della Fonte amministrativa, ovvero la generazione dei dati amministrativi per l'applicazione di un Regolamento amministrativo per gli Enti della Pubblica amministrazione, o per scopi gestionali. Da questo punto in poi, il ciclo continua all'interno dell'Istat. Nella seconda colonna sono rappresentate le attività del settore che si occupa della gestione dei dati amministrativi a supporto della produzione statistica: la fase di acquisizione delle forniture dei dati, il trattamento e la predisposizione degli accessi per gli scopi statistici, mentre la terza colonna mostra le attività dei settori di produzione che utilizzano i dati amministrativi per gli specifici obiettivi definiti nell'ambito del Psn che assolve ai fabbisogni statistici del Paese.

Nella QRCA sono presenti misure relative, principalmente, alle attività inerenti alla colonna centrale della figura, indicatori di qualità relativi agli altri due comparti potranno arricchire in futuro l'offerta della QRCA.

Figura 1.2 - Ciclo di vita dei dati amministrativi (DA) utilizzati a fini statistici



Con lo scopo di descrivere le attività connesse alla gestione dei dati amministrativi in Istat e documentate nella QRCA, viene utilizzato il GSBPM, ovvero il *Generic statistical business process model*, un framework internazionale che descrive, attraverso una terminologia armonizzata, la sequenza delle attività che vengono effettuate nel processo di produzione della statistica ufficiale[2] (Unece, 2019).

Il modello ha una struttura flessibile suddivisa in 2 livelli: il Livello 1 comprende 8 fasi e in ciascuna, al livello 2, sono elencati i corrispondenti sottoprocessi. Il GSBPM considera, inoltre, i processi trasversali (*Overarching processes*) che si esplicano lungo le varie fasi.

In questo contesto vengono considerate le fasi e sottofasi con riferimento alle attività che riguardano la gestione centralizzata dei dati amministrativi, escludendo le attività dei vari processi di produzione che conducono alla produzione finale dell'output. Nello schema di Figura 1.3 sono riprodotti i due livelli del GSBM e si possono visualizzare le fasi interessate: evidenziate da un ovale tratteggiato, le attività a cui il settore dei dati amministrativi collabora con gli altri settori dell'Istat, mentre con un ovale a linea continua sono evidenziate le attività di diretta responsabilità. Sono incluse le attività trasversali di *Metadata management* per quanto riguarda la gestione dei metadati amministrativi (cioè relativi ai dati amministrativi) e la gestione dei metadati di processo e di *Quality management* esplicata dalla QRCA a presidio delle varie funzioni.

L'individuazione dei fabbisogni informativi *Specify Needs* è la fase iniziale del modello. La prima sottofase coinvolta è la 1.1. *Identify data needs*. Seguendo Unece, 2019, questa sottofase comprende l'analisi iniziale e l'identificazione di quali statistiche sono necessarie e di ciò che è necessario per le statistiche. Può essere innescato da una nuova richiesta di informazioni o da un cambiamento ambientale come un budget ridotto. La pertinenza di questa sottofase deriva dalla seguente considerazione: l'esplorazione di nuove fonti amministrative può stimolare e orientare la produzione di nuove statistiche laddove si rendono

2 *GSBPM* https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

disponibili nuovi dati che possono soddisfare nuovi bisogni informativi prima non immaginabili. Ad esempio, alcune Fonti amministrative forniscono informazioni integrate tra unità/popolazioni di diverso tipo: dati definititi Leed (*Linked Employer-Employee Data*), derivanti dalle fonti contributive che attraverso le posizioni lavorative connettono i lavoratori alle imprese, permettono di definire dei data set con attributi integrabili su entrambi le tipologie di unità; relazioni tra gli studenti e le scuole/università presso cui svolgono i loro percorsi di studio nei dati amministrativi dell'Istruzione. In questi casi le potenzialità di uso sono notevoli grazie all'elevato dettaglio dei dati, difficilmente rilevabile attraverso un'indagine, e alla possibilità di effettuare analisi di tipo longitudinale. Il supporto a questa sottofase si esplica con la gestione delle relazioni con gli Enti titolari dei dati amministrativi coordinata dalla struttura centralizzata al fine di esplorare nuove fonti o nuove variabili non ancora acquisite. Inoltre, la produzione della documentazione dei dati amministrativi attraverso la QRCA permette agli utenti di esplorare scenari di produzione innovativi rispetto alle nuove esigenze informative o affrontare necessità di riduzioni di budget.

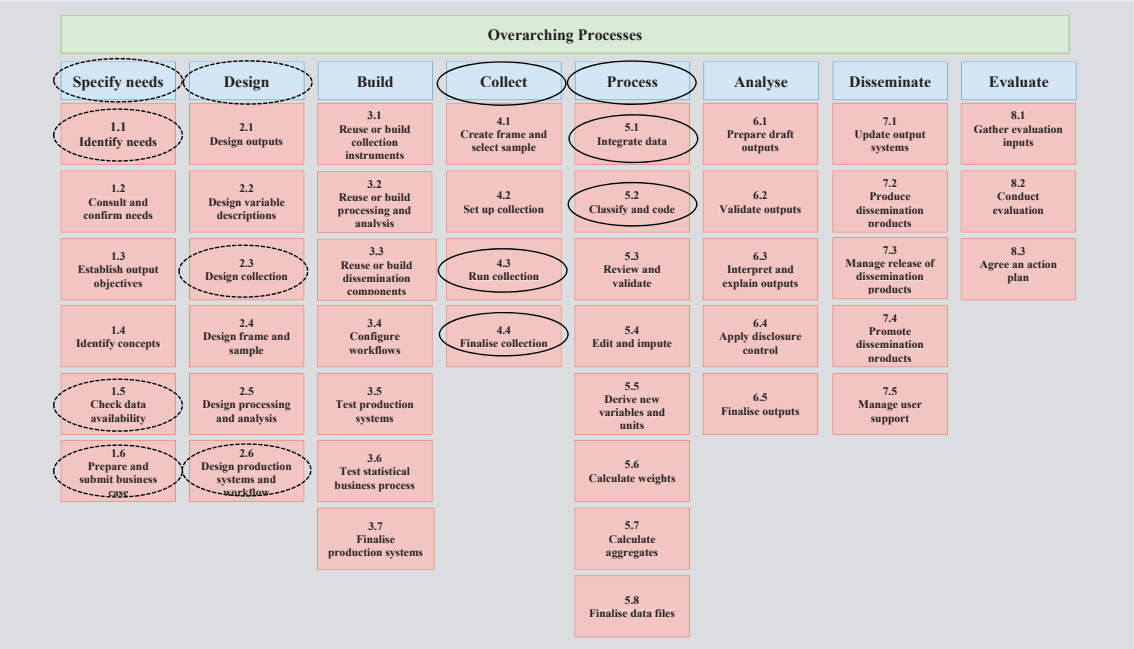
La sottofase 1.5. *Check data availability*, nell'ottica specifica della gestione dei dati amministrativi, prevede una valutazione della disponibilità ed usabilità delle fonti di dati amministrativi esistenti al fine di soddisfare i fabbisogni statistici identificati. A questo scopo il portale della QRCA può supportare i processi di produzione che intendono inserire innovazioni: le informazioni sui dati amministrativi già acquisiti dall'Istat ed utilizzati da altri processi in termini di tipi di unità, variabili, classificazioni presenti, di qualità, puntualità e tempestività dei dati possono contribuire alla valutazione della loro usabilità. Per le nuove fonti da acquisire vale la funzione di coordinamento delle relazioni con gli Enti che supporta le attività di esplorazione e di valutazione degli aspetti tecnici.

Un ulteriore supporto viene fornito per il Business case (1.6 *Prepare and submit the Business case*), sempre nell'ottica dei dati amministrativi, con lo scopo di verificarne l'accessibilità, ovvero di valutare la necessità di stipulare una convenzione o accordo per lo scambio dei dati, considerare eventuali costi di acquisizione, analizzare gli aspetti tecnici dell'acquisizione. Per quanto riguarda, invece, gli Archivi già acquisiti dall'Istat, in questa sottofase si tratta di valutare: la possibile modifica delle richieste in considerazione dei nuovi fabbisogni; il possibile inserimento di nuove variabili da acquisire; l'eventuale modifica della periodicità, della tempestività o la possibilità di acquisire eventuali dati preliminari ad anticipazione dei definitivi.

In generale, le attività della Fase di Specificazione dei fabbisogni (*Specify needs*), sono supportate annualmente dalla predisposizione del Programma annuale di acquisizione delle forniture di dati amministrativi. In prossimità del nuovo anno, il settore che si occupa dell'acquisizione dei dati amministrativi effettua una ricognizione delle esigenze da parte dei settori di produzione. Si tratta di confermare o meno le acquisizione programmate nell'anno precedente, rilevare nuovi fabbisogni sia rispetto all'acquisizione di nuovi archivi che rispetto a mutate esigenze dal punto di vista dell'usabilità degli archivi già acquisiti in termini di contenuti informativi e tempestività. La stesura del Programma si conclude con la stipula degli accordi con gli Enti titolari delle Fonti amministrative.

Nella Fase di progettazione (*Design*), vengono definiti gli strumenti di raccolta dei dati (*Design collection*), mentre nella fase di costruzione questi vengono messi in opera (*Reuse and built collection instrument*). Gli IT Tool utilizzati per l'acquisizione ed il trattamento dei dati amministrativi sono: Arcam e SIM (Sistema Integrato di Microdati), oltre alla QRCA stessa che supporta i processi.

Figura 1.3 - Le fasi di acquisizione e pretrattamento dei dati amministrativi in Istat nell'ambito del GSBPM



Il Sistema Arcam è un applicativo per l'acquisizione via web dei dati amministrativi, i suoi obiettivi sono:

- ottimizzare il processo di acquisizione dei dati amministrativi in conformità con gli standard di sicurezza IT;
- gestire i sistemi centralizzati di archiviazione nel rispetto delle regole di protezione dei dati;
- implementare il Programma annuale di acquisizione delle forniture di dati amministrativi e supportarne la gestione del monitoraggio.

Esso è composto da quattro principali parti: il portale di acquisizione, il repository dei dati, il DB dei metadati e il sistema di rilascio delle forniture che non contengono dati personali (Drovandi *et al.*, 2017).

Il fornitore dei dati amministrativi accede al portale attraverso le sue credenziali e può effettuare l'upload delle forniture pianificate nel Programma annuale. Gli amministratori, attraverso il sistema di gestione (*backoffice*), possono gestire le forniture e monitorare lo stato dell'acquisizione.

Il Sistema Integrato di Microdati (SIM) nasce nel 2013 come una struttura informativa di base comprendente l'integrazione concettuale e fisica dei microdati acquisiti da fonti amministrative con lo scopo di supportare trasversalmente i processi di produzione statistica dell'Istat (Ambroselli, 2015). In particolare, l'obiettivo è di gestire la parte del processo statistico di individuazione delle unità statistiche di base contenute nei dati amministrativi: gli Individui, le Unità economiche (imprese, istituzioni, enti pubblici) e i Luoghi (indirizzi) con lo scopo di rendere possibile l'integrazione dei dati amministrativi e sfruttarne a pieno le potenzialità per i fini statistici. In seguito, il funzionamento estremamente parametrizzato del SIM è stato adattato per garantire la procedure di pseudonimizzazione dei dati personali nel rispetto della normativa vigente (Reg. UE 2016/678).

Proseguendo nell'ambito del GSBPM, la fase vera e propria di raccolta comprende le sottofasi di *Run* e *Finalize collection*. Il processo di acquisizione dei dati prende avvio

dalla redazione del Programma Statistico Nazionale (Psn): i referenti dei lavori Istat devono dichiarare quali sono gli Archivi amministrativi che intenderanno utilizzare nel proprio processo di produzione. Questa dichiarazione costituisce il primo passo per l'acquisizione e l'accesso ai dati amministrativi. Sulla base del Programma annuale di acquisizione delle forniture di dati amministrativi, si predispongono le lettere di richiesta ufficiale dei dati agli Enti titolari e si procede all'acquisizione. La finalizzazione dell'acquisizione consiste nella verifica della validità dei dati forniti, passando per l'analisi concettuale Entità/Relazioni dei dati, i processi ETL e i controlli di conformità e completezza delle forniture. Così come le procedure di ETL, anche la fase di trattamento dei dati per l'integrazione (*Process - Integrate data*) è affidata al SIM che rende possibile, per i processi di produzione, integrare i dati utilizzando come chiave di linkage degli appositi codici pseudonimi assegnati agli Individui e alle Unità economiche nel rispetto della normativa per il trattamento dei dati personali (Regolamento Europeo 2016/678). La documentazione del processo di pseudonimizzazione delle unità statistiche è descritto nel paragrafo 2.4.4.

Ancora relativa della fase 5. *Process*, la sottofase di *Classify and code* riguarda sia la gestione delle classificazioni amministrative, sia la codifica (pseudonimizzazione) delle unità, successiva all'attività di integrazione, entrambe effettuate nell'ambito del SIM.

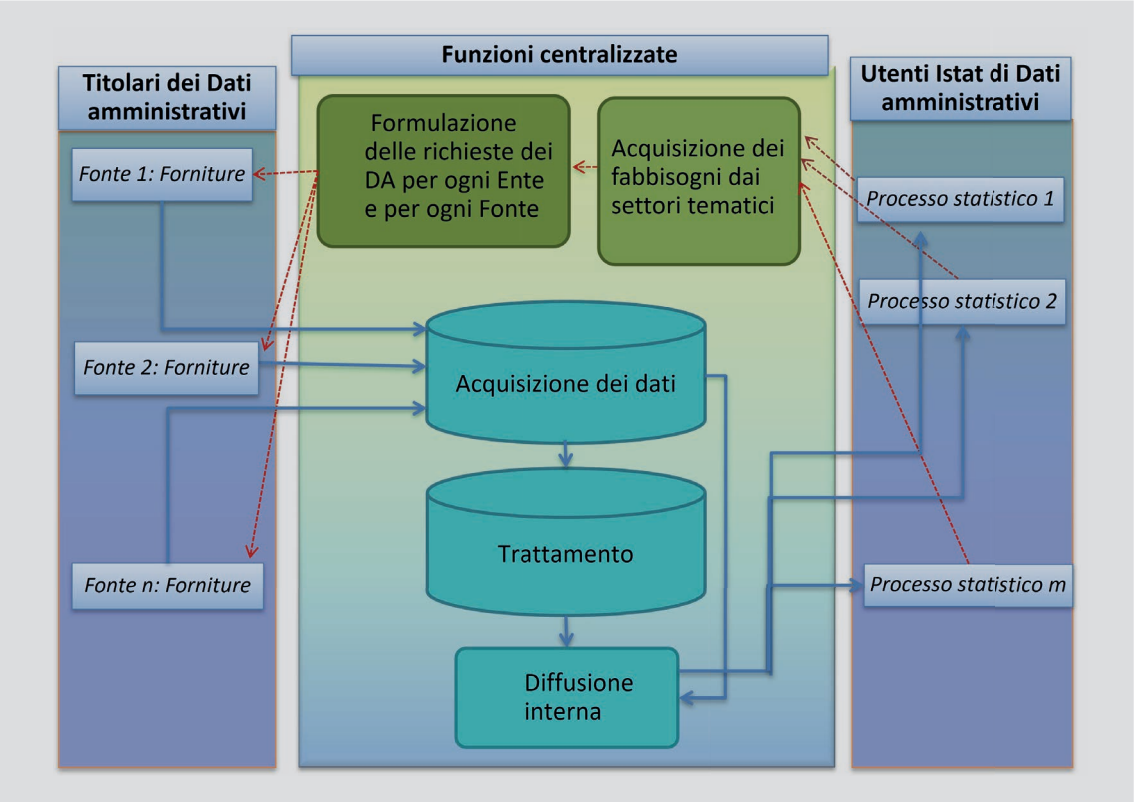
Le funzioni svolte centralmente si concludono con la messa a disposizione dei dati amministrativi per i processi di produzione che ne abbiano fatto richiesta; tale operazione, quando coinvolge Dati personali, è parte di una specifica procedura definita appositamente per garantire il rispetto della normativa (Regolamento Europeo 2016/679)[3].

Infine, per quanto riguarda gli *Overarching processes*, è nella fase trasversale di gestione della qualità (*Quality management*) che si valuta e documenta la qualità dei dati amministrativi acquisiti dall'Istat e lo strumento adottato in Istat è proprio la QRCA descritta in questo volume. Un'altra funzione trasversale riguarda la gestione dei metadati (*Metadata management*). Non sempre i dati amministrativi possiedono un corredo esaustivo di metadati, in ogni caso, i metadati forniti dall'Ente titolare della Fonte amministrativa vengono sistematizzati ed utilizzati per i processi di acquisizione e di trattamento dei dati e vengono messi a disposizione nella QRCA: si tratta dei metadati strutturali o descrittivi delle unità e delle variabili e di altra documentazione messa a disposizione dagli Enti. Negli ultimi tempi, acquisisce sempre più importanza la funzione di protezione dei dati personali ascrivibile alla fase trasversale di *Data management* (Unece, 2019). Tale funzione deve applicarsi sin dalle prime fasi del processo di produzione e, per i dati amministrativi – come già evidenziato – un ruolo importante è svolto dalla struttura centralizzata di acquisizione e pretrattamento.

Nella seguente figura è riportato un riepilogo del ciclo di vita dei dati amministrativi: i soggetti coinvolti si coordinano per definire i fabbisogni in relazioni alle disponibilità (frecce rosse tratteggiate); da questo scambio informativo si origina il flusso dei dati (frecce azzurre continue).

3 Questa attività potrebbe essere rappresentata nel GSBPM operando un riadattamento delle fasi relative all'output statistico (6. Analyse, 7. Disseminate) all'output intermedio generato dal processo di gestione centralizzata dei dati amministrativi.

Figura 1.4 - Il processo di acquisizione ed il flusso dei dati amministrativi



1.3 Il framework della qualità adottato

Al fine di inquadrare i concetti della qualità statistica dei dati amministrativi e della loro usabilità nei processi di produzione, in Istat è stato adottato un framework teorico che prevede un approccio gerarchico multidimensionale che permette di classificare le informazioni in modo flessibile e chiaro. Esso comprende delle Iperdimensioni della qualità denominate FONTE, METADATI E DATI; all'interno di ciascuna sono presenti le corrispondenti Dimensioni della qualità. A loro volta le Dimensioni sono descritte da Indicatori applicati con specifici Metodi di misura adattabili ai diversi contesti.

Il framework Istat è basato sull'idea originariamente definita da *Statistics Netherlands* e poi sviluppata nell'ambito del progetto internazionale BLUE ETS, WP4, le misure della qualità sono state successivamente definite in base al contesto dell'Istat (Daas *et al.*, 2009; Daas *et al.*, 2011c; Cerroni *et al.*, 2014). Nel Prospetto 1.1 sono sintetizzate le Dimensioni della qualità dei dati amministrativi. La documentazione ed i report della QRCA sono organizzati secondo questo framework.

L'Iperdimensione FONTE riporta le informazioni della Fonte dei dati amministrativi e i principali attributi dell'Archivio derivato dalla Fonte. In dettaglio, le Dimensioni della qualità comprendono: le informazioni anagrafiche di base, la Rilevanza e gli usi statistici, le questioni connesse alle norme di accesso ed uso dei dati nel rispetto della normativa sul trattamento dei dati personali e, infine, una descrizione degli accordi che regolano gli scambi dei dati con il titolare della Fonte, le relazioni e i feedback in caso di problematiche e la relativa documentazione (Convenzioni, Gruppi di lavoro inter-istituzionali, …).

Prospetto 1.1 - Il framework della qualità dei dati amministrativi adottato in Istat

IPERDIMENSIONE	DIMENSIONE
FONTE Informazioni necessarie a gestire il processo di acquisizione dei dati con lo scopo di valutare e migliorare la qualità dei dati acquisiti	Informazioni di base
	Rilevanza e usi statistici
	Riservatezza e protezione dei dati
	Accordi con l'Ente titolare, relazioni e feedback
METADATI Informazioni per la valutazione della qualità a livello concettuale e di processo	Chiarezza e interpretabilità
	Comparabilità tra concetti amministrativi e concetti statistici
	Stabilità temporale dei concetti amministrativi
	Descrizione del processo di acquisizione e trattamento da parte del titolare
DATI Valutazione della qualità dei dati acquisiti	Stato delle forniture
	Aspetti temporali
	Controlli tecnici
	Integrabilità/Integrazione
	Accuratezza e coerenza interna
	Completezza

L'Iperdimensione METADATI, declinata per le unità e per gli oggetti (eventi, variabili) nella Dimensione della *Chiarezza/Interpretabilità* contiene la descrizione del contenuto informativo dell'Archivio: la tipologia degli oggetti e l'elenco delle variabili con le corrispondenti classificazioni amministrative disponibili per le variabili categoriche. La dimensione della *Comparabilità* concettuale riguarda il confronto tra i concetti amministrativi e i concetti statistici con lo scopo di misurarne la distanza.

La dimensione della *Stabilità temporale dei concetti amministrativi* considera la necessità di documentare i cambiamenti che possono modificare l'usabilità statistica dei dati a causa di modifiche normative o gestionali o per variazioni strutturali dei dati connesse alla Fonte.

L'ultima dimensione della qualità nell'ambito dell'Iperdimensione METADATI concerne i possibili trattamenti sui dati effettuati dall'Ente: in questo caso è bene acquisire tutte le informazioni disponibili al fine di poter correttamente utilizzare i dati per i fini statistici e documentarne la qualità.

L'Iperdimensione dei DATI comprende la valutazione stessa dei dati acquisiti. La prima dimensione relativa allo *Stato delle forniture* comprende l'elenco delle forniture dei dati con le corrispondenti caratteristiche definitorie ed il monitoraggio del processo di acquisizione e pre-trattamento centralizzato. Nella Dimensione degli *Aspetti temporali* sono presenti gli Indicatori di Puntualità (rispetto delle scadenze nella consegna dei dati da parte dell'Ente fornitore) e di Tempestività (distanza tra la data di arrivo della Fornitura in Istat e l'ultima data degli eventi registrati nel dataset); si possono fornire, inoltre, informazioni sulla dinamica degli oggetti e sulla stabilità delle variabili da restituire in serie storica. La terza Dimensione riguarda la fase dei cosiddetti *Technical checks*, o *Controlli tecnici*: essi hanno lo scopo di verificare la conformità dei dati ricevuti, rispetto ai dati attesi in seguito alla richiesta ufficiale da parte dell'Istat; l'esito di tali controlli è una prima forma di validazione della fornitura dei dati, in caso si riscontrino dei problemi si procede, il più celermente possibile, a ricontattare l'Ente per chiedere informazioni o, quando necessario, a rinviare i dati.

È molto importante che un Archivio sia integrabile con altri dataset e questo è possibile laddove ci siano delle variabili di linkage sufficientemente estese e di buona qualità. Nella Dimensione *Integrabilità/Integrazione* vengono considerate alcune misure relative alle po-

tenzialità di integrazione che documentano la presenza e la qualità delle variabili di linkage utilizzate, sostanzialmente, per l'identificazione delle unità statistiche all'interno dei dataset amministrativi; la dimensione dell'Integrazione riporta indicatori della qualità del record linkage.

L'*Accuratezza* ha l'obiettivo di misurare l'inconsistenza dei dati per le unità, per le relazioni, per le variabili e loro combinazione.

Nella dimensione della *Completezza* dei dati si misurano due aspetti: rispetto alle unità sono previsti indicatori di copertura; rispetto alle variabili la percentuale dei valori mancanti.

Nel prossimo Capitolo viene descritto il portale della QRCA scendendo nel dettaglio delle dimensioni, degli indicatori della qualità e delle misure adottate.

1.4 Tipi di utilizzo della QRCA

La QRCA nasce per rispondere alle esigenze emerse dai settori tematici di produzione della statistica ufficiale ed è accessibile solo per gli utenti interni all'Istat attraverso l'ambiente intranet dell'Istituto. Nel tempo la sua utilità si è estesa e attualmente svolge diverse funzioni (Figura 1.5).

Per i settori di produzione, la QRCA è utile sia per coloro che utilizzano ormai a regime i dati amministrativi come input dei processi, sia per coloro che intendono esplorare le potenzialità dei dati già acquisiti dall'Istat ed innovare i processi di produzione di cui sono responsabili.

Per i primi, la funzione di monitoraggio delle acquisizioni, disponibile nell'Iperdimensione dei DATI – report Stato delle forniture, permette di verificare in tempo reale la disponibilità dei dati in Istituto a valle del processo di acquisizione e dell'eventuale pretrattamento. La seconda funzione rivolta agli utenti dei dati amministrativi è il monitoraggio della qualità dei dati effettuata attraverso gli specifici Controlli tecnici (§ 2.4.3): essa permette di evidenziare possibili discontinuità del livello della qualità dei dati e, in caso di problemi rilevanti, rende possibile agire tempestivamente e ricontattare l'Ente per ricevere spiegazioni sulle possibili cause non comunicate preventivamente. Se si riscontrano errori tecnici di estrazione dei dati o problemi di invio, si procede nella richiesta di una nuova fornitura di dati all'Ente.

In generale, la produzione della documentazione della QRCA, riferita ai dati amministrativi utilizzati come input del processo, può integrare la reportistica di qualità dell'output.

La funzione di usabilità dei dati a scopi statistici è ormai consolidata per l'Istituto: in prima analisi, è possibile verificare – per ciascuna Fonte e per ciascun Archivio amministrativo – se i dati sono idonei a supportare la produzione rispetto ai contenuti informativi, alla tempestività, alle classificazioni presenti e così via.

Un'altra funzione svolta dalla QRCA è di fornire un supporto di documentazione al personale incaricato dell'acquisizione e del trattamento dei dati. La conservazione delle informazioni e degli indicatori in serie storica dal 2017 e il tempestivo aggiornamento della QRCA permettono di disporre di un ausilio sia per rispondere ai fabbisogni degli utenti, sia per gestione delle operazioni correnti. La QRCA, inoltre, fornisce la reportistica per la rendicontazione delle attività svolte.

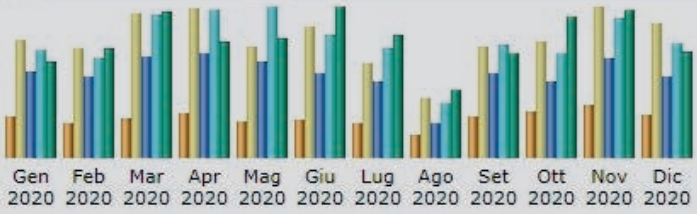
Infine, per il coordinamento delle relazioni con gli Enti titolari dei dati amministrativi, si possono definire dei report ad hoc, da valutare caso per caso, utili per rafforzare il coinvolgimento ed evidenziare possibili miglioramenti della qualità statistica dei dati.

Figura 1.5 - Le funzioni della QRCA

Per concludere questo paragrafo, in Figura 1.6 si presentano alcune statistiche di accesso al sito intranet della QRCA relative all'anno 2020. Le visite mensili sono circa 300 (in media 294 visite al mese) per circa 100 utenti diversi (97 in media nel 2020), le pagine visitate in un mese sono in media 4.650. Nel primo semestre del 2021 si confermano, con un leggero aumento, i numeri del 2020 decretando la reale utilità dello strumento.

Figura 1.6 - Le statistiche di accesso al portale della QRCA. Anno 2020

Riepilogo mensile



Mese	Visitatori diversi	Numero di visite	Pagine	Accessi	Banda usata
Gen 2020	102	289	4.790	5.977	90.72 MB
Feb 2020	83	269	4.495	5.526	102.73 MB
Mar 2020	95	357	5.646	7.974	137.67 MB
Apr 2020	109	370	5.826	8.193	109.71 MB
Mag 2020	88	274	5.331	8.345	112.62 MB
Giu 2020	92	325	4.647	6.861	141.42 MB
Lug 2020	86	231	4.231	6.099	114.77 MB
Ago 2020	56	146	1.879	3.045	64.23 MB
Set 2020	101	275	4.679	6.303	98.54 MB
Ott 2020	115	285	4.205	5.762	132.94 MB
Nov 2020	131	371	5.542	7.750	139.68 MB
Dic 2020	106	331	4.528	6.388	99.27 MB
Totale	1.164	3.523	55.799	78.223	1.31 GB

Fonte: AWStats - Statistiche del sito qrca.istat.it

2. I CONTENUTI INFORMATIVI DELLA QRCA

Il portale della QRCA è stato pubblicato in Istat per la prima volta il 9 novembre 2018 ed è disponibile all'indirizzo http://qrca.istat.it agli utenti Istat nell'ambiente intranet.

La QRCA si alimenta in modo automatico riutilizzando i metadati derivanti dai processi di acquisizione, trattamento e gestione dei dati amministrativi, come descritto più in dettaglio nel Capitolo 3. Il suo processo di implementazione è stato graduale: nella prima versione prototipale sono state pubblicate le informazioni derivabili dai metadati allora disponibili, più precisamente 38 misure. Altre 11 sono state aggiunte nella Versione 2.0, pubblicata ad aprile 2021, che ha avuto lo scopo di avviare il primo passo della fase di adeguamento della documentazione alle indicazioni fornite dal Garante per la protezione dei dati personali in merito al trattamento dei dati amministrativi (Garante sulla protezione dei dati personali, 2020).

Delle 106 misure definite teoricamente e riportate in Appendice, ne rimangono da implementare 57, esse si articolano secondo più criteri di implementabilità derivanti da vincoli tecnici di calcolo, alimentazione del sistema e, non ultimo, il rapporto costo/beneficio connesso. Nel corso del Capitolo 2 si entra nel dettaglio delle varie misure, alla fine del volume sono riportate alcune valutazioni sugli sviluppi futuri.

Si sottolinea che la QRCA non tratta microdati, ovvero dati personali, ma solo metadati e macrodati; quindi, non sono necessarie specifiche misure tecniche di protezione dei dati.

2.1 L'anagrafe degli archivi e le informazioni disponibili

La QRCA si presenta con una Home page che invita all'autenticazione attraverso le credenziali dell'Istituto. Una volta avvenuto l'accesso, si effettua la selezione dell'Archivio di interesse mediante l'ausilio di una funzione di ricerca che prevede l'inserimento di parole del nome del titolare o del nome dell'Archivio. Successivamente è possibile navigare tra i report disponibili organizzati secondo il framework della qualità.

La lista degli archivi per i quali è disponibile la documentazione comprende tutti gli archivi acquisiti dal Servizio responsabile presso la Direzione preposta alla raccolta dati dell'Istat. Poiché il sistema della QRCA si alimenta automaticamente con gli IT Tool di gestione dei dati amministrativi ed il sistema centralizzato di acquisizione è entrato a regime nel 2017, sono documentati gli archivi acquisiti dal 2017. Sono quindi esclusi gli archivi acquisiti in precedenza e non più acquisiti dal 2017. Per costruzione, sono inoltre esclusi gli archivi amministrativi utilizzati in Istat ma che non sono acquisiti centralmente dal Servizio.

Una volta selezionato l'Archivio, si accede al primo report dell'Iperdimensione FONTE e un menu permette la navigazione tra i vari report che compongono la QRCA: le tre Iperdimensioni FONTE, METADATI e DATI e i livelli gerarchici successivi relativi alle Dimensioni della qualità. La mappa della navigazione è riportata schematicamente nel Prospetto 2.1.

Nell'attuale versione, sono presenti 2 report per l'Iperdimensione FONTE, 8 report per METADATI e 9 report per DATI; completano la documentazione 3 ulteriori report riassuntivi accessibili direttamente senza la preliminare selezione dell'Archivio: Monitoraggio delle acquisizioni, Tabelle di sintesi dei processi di acquisizione e Tabelle di sintesi dei processi di

trattamento, per un totale di 22 report. Ciascun report può essere composto da più tabelle e grafici.

Prospetto 2.1 - Mappa della navigazione nel sito della QRCA

FONTE	METADATI			DATI	
Informazioni di base	Chiarezza e interpretabilità dei dati	Oggetti amministrativi		Stato delle forniture	
Rilevanza ed usi statistici		Variabili amministrative	Dati forniti dall'Ente	Aspetti temporali	Tempestività e Puntualità
Riservatezza e protezione dei dati			Viste per gli utenti SIM	Controlli tecnici -> Conformità	Numero record
Accordi con l'Ente titolare, relazioni e feedback			Tabelle di SIM		Campi valorizzati
			Strutture dei dati in SIM		Frequenze
		Classificazioni amministrative	Classificazioni fornite dall'Ente		Decodifiche mancanti
			Connessioni tra classificazioni	Integrabilità/ Integrazione dei dati	Linkabilità delle unità
		Altra documentazione disponibile			Monitoraggio del Processo di integrazione
	Comparabilità tra concetti amministrativi e concetti statistici (unità e variabili)			Completezza delle variabili	
	Descrizione del processo di acquisizione e trattamento (da parte del titolare)			Completezza delle unità	
				Accuratezza e coerenza interna	

RAPPORTI RIASSUNTIVI
Monitoraggio delle acquisizioni
Tabelle di sintesi dei processi di acquisizione
Tabelle di sintesi dei processi di trattamento

A causa della strategia di produzione della QRCA (vedi Capitolo 3), non tutte le informazioni sono disponibili per tutti gli Archivi. Se un Archivio non viene trattato e gli utenti accedono ai dati così come vengono consegnati dall'Ente, non sono disponibili i metadati di processo che alimentano il calcolo degli indicatori.

Le disponibilità dei report, quindi, dipendono dal livello di trattamento a cui è sottoposto l'Archivio: nel seguente prospetto vengono riassunti i report disponibili per ciascuna Iperdimensione e Dimensione (la disponibilità dei report è segnalata con il simbolo 'x'. Nelle prossime versioni si cercherà di ampliare le informazioni disponibili.

Ogni report visualizzato può essere utilmente esportato in formato Excel o in formato PDF per effettuare delle elaborazioni o produrre documentazione sul processo che utilizza i dati dell'Archivio oggetto di analisi.

Nei paragrafi successivi si presenta il contenuto di ciascun report.

Prospetto 2.2 - Report della QRCA disponibili per ciascuna Iperdimensione della qualità per Tipologia di trattamento dell'Archivio

IPERDIMENSIONE	Report	Tipologia di trattamento			
		Nessun trattamento in SIM	Trattamento in SIM senza assegnazione dei codici sim	Trattamento in SIM con assegnazione del codice sim a livello di individuo	Trattamento in SIM con assegnazione del codice sim a livello di unità economica
FONTE	Informazioni di base	x	x	x	x
	Rilevanza e Usi statistici	x	x	x	x
METADATI	Chiarezza/Interpretabilità -> Oggetti Amministrativi		x	x	x
	Chiarezza/Interpretabilità -> Variabili amministrative		x	x	x
	Chiarezza/Interpretabilità -> Classificazioni amministrative		x	x	x
	Chiarezza/Interpretabilità -> Altra documentazione disponibile	x	x	x	x
DATI	Stato delle forniture	x	x	x	x
	Aspetti temporali – Tempestività e Puntualità	x	x	x	x
	Controlli tecnici -> Conformità		x	x	x
	Integrabilità/Integrazione dei dati			x	
RAPPORTI RIASSUNTIVI	Monitoraggio Tabelle di Sintesi	Report accessibili senza la selezione dell'Archivio			

2.2 L'Iperdimensione FONTE

La Fonte amministrativa, come riportato nel paragrafo 1.1.1, è "Un insieme di dati raccolti e mantenuti per l'attuazione di uno o più regolamenti amministrativi. In un senso più ampio, qualsiasi fonte di dati che contiene informazioni che non sono raccolte principalmente per scopi statistici". Costituisce, quindi, l'insieme dei dati gestiti dal titolare. L'Istat, per gli scopi statistici di produzione della statistica ufficiale e per le finalità della ricerca, acquisisce sistematicamente – ad intervalli temporali predefiniti – delle porzioni dei dati della Fonte: l'insieme dei dati via via acquisiti, definisce il concetto di Archivio amministrativo.

Nell'Iperdimensione FONTE, vengono considerate tutte le misure riferite a questi due concetti. In genere non si hanno molte informazioni sulla Fonte in sé e la maggior parte dell'informazione disponibile è riferita all'Archivio amministrativo derivato. Occorre specificare che da una stessa Fonte amministrativa possono derivare più Archivi amministrativi.

Nei prossimi paragrafi si descrivono le misure presenti nelle Dimensioni della Fonte:

- Informazioni di base;
- Rilevanza e usi statistici;
- Riservatezza e protezione dei dati;
- Accordi con l'Ente titolare, relazioni e feedback.

2.2.1 Informazioni di base

Per ogni Archivio amministrativo acquisito dall'Istat, il primo Report della QRCA, denominato *FONTE->Informazioni di base* mostra le seguenti misure (Prospetto 2.3).

Prospetto 2.3 - Indicatori, informazioni/misure del report FONTE - Informazioni di base

INDICATORE	Informazione/Misura
Identificazione della Fonte	Denominazione Archivio (contiene il nome della Fonte)
Identificazione della Fonte	Denominazione dell'Ente titolare della Fonte amministrativa
Riservatezza e protezione dei dati	Presenza di dati personali (Reg. UE 2016/678) (SI/NO)
Riservatezza e protezione dei dati	Presenza di dati rientranti in particolari categorie (ex sensibili) (SI/NO)
Riservatezza e protezione dei dati	Presenza di dati relativi a condanne penali e reati (ex guidiziari) (SI/NO)
Lunghezza della serie dei dati	Serie storica dei dati disponibili (anni disponibili)
Tipologia di trattamento	Tipologia di trattamento a cui è sottoposto l'Archivio
Lunghezza della serie dei dati	Serie storica disponibile nel SIM (anni)
Tipologia di trattamento	Codifica degli indirizzi in RSBL (SI/NO)
Lunghezza della serie dei dati	Anni in RSBL
Gestione dell'acquisizione	Referente Istat per l'acquisizione

Oltre alle variabili anagrafiche, l'attributo principale della Fonte è il titolare della Fonte amministrativa da cui l'Archivio è acquisito che è il referente Istat per il processo di acquisizione dei dati e dei metadati descrittivi, se disponibili.

Nell'ambito delle Informazioni di base sono comprese alcune indicazioni sull' Accessibilità/Riservatezza, ovvero le norme che permettono l'accesso e l'uso dei Dati Personali eventualmente contenuti nell'Archivio. In particolare, si riporta se l'Archivio comprende dati personali secondo la definizione dal Reg. UE 2016/679 - articolo 4:

«dato personale»: qualsiasi informazione riguardante una persona fisica identificata o identificabile («interessato»); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale.

Nel caso di presenza di Dati personali si specifica, inoltre l'eventuale presenza di dati rientranti in particolari categorie (ex sensibili), ovvero "dati personali che rivelino l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche, o l'appartenenza sindacale, nonché … dati genetici, dati biometrici intesi a identificare in modo univoco una persona fisica, dati relativi alla salute o alla vita sessuale o all'orientamento sessuale della persona" (Reg. UE 2016/679, art. 9), e di dati relativi a condanne penali e reati (ex giudiziari), (Reg. UE 2016/679, art. 10).

Queste informazioni sono determinanti per le modalità di accesso ai dati; infatti, nel caso di presenza di Dati personali diventa vincolante l'applicazione di una specifica procedura in cui occorre definire le finalità statistiche di uso dei dati, la delimitazione dei dataset da utilizzare e le autorizzazioni per le singole utenze personali di accesso.

Nella QRCA, laddove possibile, gli elementi caratteristici dell'accesso ai dati sono documentati, come si vedrà in seguito, nei vari report.

L'informazione della Lunghezza della serie storica dei dati si riferisce alla serie delle annualità disponibili in Istat, ovvero per le quali è stato completato il processo di acquisizione e eventuale trattamento[1].

1 La serie storica degli anni disponibili pubblicata nella QRCA può essere troncata a sinistra qualora l'archivio sia stato acquisito prima dell'anno 2017 e non sia stato trattato in SIM. Questo limite dell'informazione fornita deriva dalle tempistiche del processo di informatizzazione centralizzata delle acquisizioni conclusosi nel 2017.

Un'ulteriore informazione contenuta nel report è la Tipologia di trattamento centralizzato a cui è sottoposto l'Archivio, prima del suo rilascio agli utenti Istat per gli usi statistici. Le Tipologie di trattamento cui sono sottoposti i dati amministrativi sono diverse e dipendono dalle caratteristiche dell'Archivio stesso e dalle esigenze degli utilizzatori. La decisione sulla tipologia di trattamento può cambiare nel tempo in relazione a sopraggiunti nuovi fabbisogni degli utenti.

Le tipologie di trattamento effettuate nell'ambito del SIM possono essere raggruppate nelle seguenti quattro categorie.

Nessun trattamento in SIM

Per l'Archivio è previsto il rilascio dei dati grezzi, così come sono forniti dall'Ente, agli utenti interni autorizzati. In generale, gli archivi non trattati da SIM sono quelli privi di microdati riferiti ad unità statistiche target (Individui, Unità economiche).

Trattamento in SIM senza assegnazione dei Codici SIM

I dati delle forniture dell'Archivio vengono caricati nel DB SIM secondo le regole del DB relazionale, i corrispondenti metadati vengono inseriti (per le nuove acquisizioni) o aggiornati (per le forniture abituali) nel sistema dei metadati di SIM denominato SISME che comprende i cataloghi delle variabili, delle classificazioni, delle tabelle dei dati e di tutti gli oggetti necessari alla gestione. Non vengono assegnati i codici SIM. In generale, gli Archivi a cui è associato questo trattamento sono privi di microdati riferiti ad unità statistiche target (Individui, Unità economiche), il trattamento viene effettuato per rendere i dati più fruibili agli utenti.

Trattamento in SIM con assegnazione del Codice SIM a livello di individuo

Oltre al trattamento precedente, viene avviata la procedura di assegnazione, a livello di individuo, del codice SIM invariante nel tempo e tra gli archivi. Tale processo permette la pseudonimizzazione dei dati.

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

Questo trattamento è simile al precedente solo che la procedura di codifica avviene a livello di unità economica. Si procede con l'assegnazione preliminare del codice SIM Unità Giuridica e si conclude con l'assegnazione del codice Unità Economica derivato dall'Archivio Statistico delle Imprese Attive (ASIA). Il processo permette la pseudonimizzazione dei dati.

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

Questo trattamento è l'unione dei tre trattamenti precedenti e riguarda gli archivi che contengono informazioni riferite sia agli Individui che alle Unità economiche. Essi sono di particolare importanza poiché permettono di collegare i due tipi di unità, è il caso dei dati previdenziali, cosiddetti di tipo LEED (*Linked employer-employee data*) in cui le informazioni sono riferite sia ai datori di lavoro (Unità economiche) che ai lavoratori (Individui).

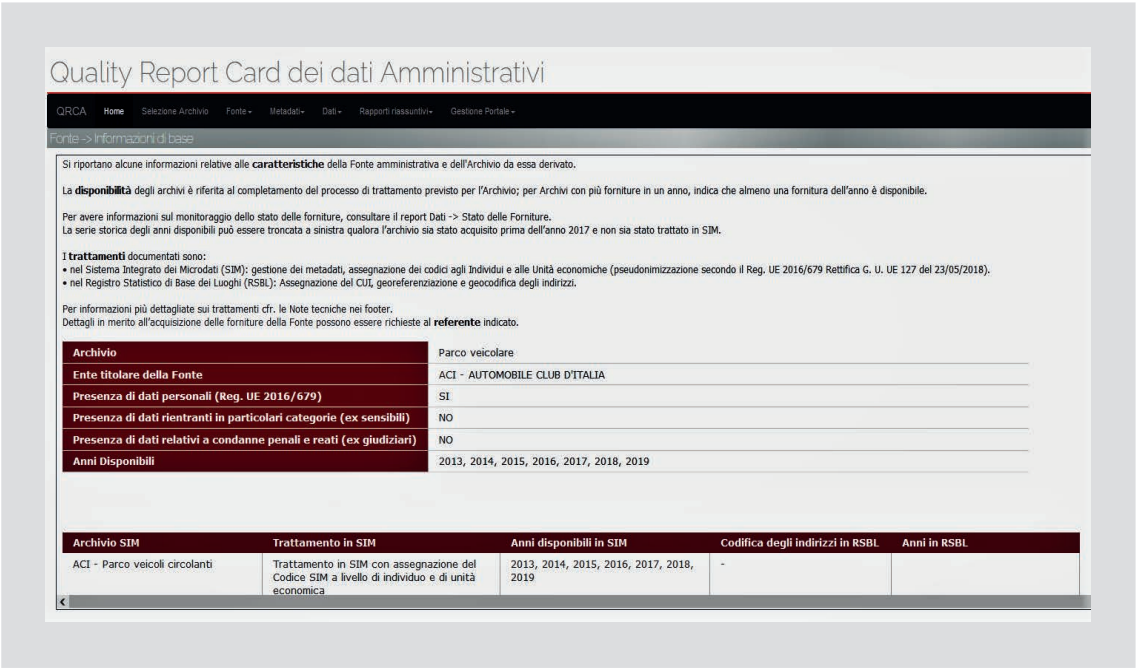
Un ulteriore trattamento che viene documentato, è l'assegnazione di codici pseudonimi agli indirizzi degli Individui e delle Unità economiche, quando presenti nell'Archivio ed utilizzati a scopi statistici. Gli indirizzi vengono trattati al fine di individuarli univocamente nel tempo e tra gli Archivi. Il trattamento consiste nel riconoscimento dell'indirizzo e nella conseguente apposizione del Codice Unico dell'Indirizzo (CUI) ad opera del Registro Statistico di Base dei Luoghi (RSBL) che, in generale, cura gli aspetti di georeferenziazione dei dati statistici a supporto dei Censimenti e del Sistema dei Registri dell'Istat di cui fa parte (Crescenzi, Lipizzi, 2020). Per gli archivi interessati, la QRCA indica quali annualità sono disponibili in RSBL.

Nell'ottica di ampliare l'usabilità dei dati e minimizzare i rischi di identificazione dei soggetti interessati, è in corso di applicazione un ulteriore trattamento che prevede l'assegnazione del codice SIM per gli Individui compresi nelle popolazioni delle unità economiche in quanto imprese senza personalità giuridica. Nella prossima versione della QRCA sarà documentato anche questo trattamento per ora riservato a due Forniture.

Conclude il report *FONTE ->Informazioni di base* il nome del Referente Istat che si occupa dell'acquisizione dell'Archivio e che l'utente può contattare per ulteriori informazioni tecniche.

La Figura 2.1 mostra la grafica del report.

Figura 2.1 - Il report FONTE-> Informazioni di base nella QRCA per l'Archivio Parco veicolare



2.2.2 Rilevanza e utilizzi statistici dei dati amministrativi

La dimensione della Rilevanza di un Archivio amministrativo è misurata dall'estensione di uso dell'Archivio all'interno dei processi Istat. Una misura semplice ed efficace è data dal Numero di lavori statistici del Psn a titolarità dell'Istat che utilizzano l'Archivio. In generale tale numero è sempre maggiore o uguale ad uno, poiché l'acquisizione deriva dal fabbisogno statistico[2]. Questa misura, disponibile in serie storica è affiancata all'elenco dei lavori. Una seconda misura presentata è il Numero di normative UE il cui adempimento dipende dai dati dell'Archivio, anche questo fornito in serie storica ed accompagnato dall'elenco delle Normative per ciascun anno (Prospetto 2.4).

Questi indicatori misurano anche il grado di dipendenza della produzione statistica dallo specifico Archivio amministrativo e, in questa prospettiva, forniscono un peso al grado di attenzione da porre sul monitoraggio della qualità delle forniture e sulla condivisione di responsabilità con l'Ente titolare della relativa Fonte.

2 I casi di valori pari a zero sono dovuti ai tempi di formalizzazione del Psn.

Prospetto 2.4 - Indicatori, informazioni/misure del report FONTE - Rilevanza ed usi statistici

INDICATORE	Informazione/Misura
Rilevanza	Numero dei lavori Istat del PSN che utilizzano i dati dell'Archivio per anno di aggiornamento del PSN (serie storica dal 2017)
Rilevanza	Elenco e codice dei lavori Istat del PSN che utilizzano i dati dell'Archivio per anno (serie storica dal 2017)
Rilevanza	Numero di normative UE il cui adempimento dipende dall'uso dell'Archivio a fini statistici per lavoro PSN per anno (serie storica dal 2017)
Rilevanza	Elenco della Normativa UE il cui adempimento dipende dall'uso dell'Archivio a fini statistici per lavoro PSN per anno (serie storica dal 2017)

Un'ultima dimensione presente nell'Iperdimensione della Fonte è relativa alle *Relazioni e feedback con l'Ente titolare della Fonte amministrativa*. Tale Dimensione non è attualmente implementata nella QRCA. In particolare si tratta di documentare gli accordi stipulati con l'Ente in forma di Protocolli d'Intesa, Gruppi di lavoro, *Gentlemen agreement*, Convenzioni e altro; inoltre, sarebbe importante riportare le informazioni relative alle procedure di feedback in caso di problemi nella consegna delle forniture, le azioni volte al miglioramento della qualità dei dati.

2.3 L'Iperdimensione METADATI

L'Iperdimensione dei METADATI assume un significato particolare per i dati amministrativi dal momento in cui il processo di produzione è esterno all'Istituto di statistica. Avere le informazioni necessarie alla corretta interpretazione dei dati è fondamentale per il loro uso statistico. In questa ottica, le relazioni con gli Enti mirano sempre a sollecitare la fornitura dei metadati di corredo. Le Dimensioni della qualità trattate in questo paragrafo sono:

- Chiarezza e interpretabilità;
- Comparabilità tra concetti amministrativi e concetti statistici;
- Stabilità temporale dei concetti amministrativi;
- Descrizione del processo di acquisizione e trattamento da parte del titolare.

2.3.1 Chiarezza e interpretabilità dei dati

Nella QRCA la dimensione della qualità attualmente implementata è la Chiarezza e interpretabilità dei dati ed in essa vengono riportate le varie informazioni disponibili dall'Ente e dai processi di trattamento.

Le informazioni si articolano principalmente nei tre elementi: unità amministrative, variabili amministrative e, per le variabili categoriche, classificazioni amministrative. Nel Prospetto 2.5 sono riportate le misure esposte.

Prospetto 2.5 - Indicatori, informazioni/misure del report METADATI - Chiarezza e interpretabilità dei dati

INDICATORE	Informazione/Misura
Descrizione dei dati (unità)	Tipi di unità presenti nell'Archivio
Descrizione dei dati (variabili)	Elenco delle variabili dell'Archivio fornite dall'Ente e loro caratteristiche
Descrizione dei dati (variabili)	Elenco strutturato delle Viste di SIM per i rilasci interni e delle variabili in esse contenute
Descrizione dei dati (variabili)	Elenco delle Tabelle Oracle che compongono l'Archivio e delle variabili in esse contenute
Descrizione dei dati (variabili)	Grafico delle strutture dei dati (Tabelle e chiavi Oracle)
Descrizione dei dati (variabili)	Classificazioni delle variabili categoriche presenti nell'Archivio
Descrizione dei dati	Altra documentazione disponibile

La prima misura, indica le tipologie di oggetti amministrativi presenti nell'Archivio e corrispondenti alle unità statistiche di base identificate nella fase di trattamento, ovvero gli Individui, le Unità economiche, i Luoghi. Questa caratterizzazione permette di classificare le informazioni disponibili, in futuro, ulteriori specificazioni saranno introdotte nei processi e potranno utilmente arricchire la documentazione. Ad esempio, sarebbe utile entrare nel dettaglio del tipo Individuo, se si tratta di studente, lavoratore dipendente, contribuente, ecc. e del Tipo Unità economica. Al momento le informazioni si limitano alla prima scomposizione nelle tre macro tipologie. Per i Luoghi si riporta la presenza nell'Archivio degli indirizzi trattati per l'apposizione del Codice Unico Indirizzo (CUI).

Rispetto alle unità, per le variabili amministrative si forniscono più dettagli informativi sia in relazione ai metadati forniti dall'Ente che al processo di trattamento.

Nel primo report sono mostrate le variabili che compongono la fornitura. In generale esse costituiscono un sottoinsieme delle variabili amministrative presenti nella Fonte presso l'Ente titolare, la loro selezione deriva dal processo di istruttoria che determina l'incontro tra la domanda di informazioni da parte dei processi di produzione statistica dell'Istat e la loro disponibilità.

La struttura dei dati rispecchia le caratteristiche originarie della Fonte o l'estrazione da questa derivata, quindi la presentazione dell'elenco delle variabili è associata ai vari file che compongono la fornitura: per ciascuno dei file forniti, viene pubblicato nella QRCA l'elenco delle variabili contenute. Possono verificarsi casi in cui la fornitura sia composta da un solo file e viene presentato, quindi, il relativo elenco di variabili. Per ciascuna variabile si riporta:

- Numero progressivo della variabile
- Nome della variabile nel file
- Descrizione della variabile, se disponibile
- Formato della variabile
- Separatore Decimale
- Valori Null
- Nome della classificazione, se esistente
- Nome della Tabella Oracle del DB SIM in cui viene caricata la variabile
- Eventuali Note associate alla variabile, in questo campo la sigla DS segnala la presenza di Dati Sensibili.

Nel campo delle Classificazioni, cliccando sul nome della Classificazione, si attiva un pop-up che permette di visualizzarne le varie modalità.

Il secondo report riporta l'elenco delle denominazioni dei dataset messi a disposizione degli utenti interni sotto forme di tabelle virtuali (Viste) e, per ciascuno, l'elenco delle variabili contenute. In questo caso si fornisce il nome del campo, la descrizione e l'eventuale nome della Classificazione associata. Le Viste sono suddivise per Tipo di dati in esse contenuti, in relazione alla necessità di accedere o meno ai dati personali. I tipi definiti sono:

- Dati pseudonimizzati (solo codici SIM);
- Dati con variabili identificative (senza codici SIM);
- Dati con variabili identificative e pseudonimi (con codici SIM).

Questa classificazione permette di gestire i rilasci nel rispetto della normativa vigente, senza appesantire il sistema. Nel report della QRCA si può selezionare il Tipo di dati a cui è possibile accedere in base alle autorizzazioni; conseguentemente si può visualizzare l'elenco delle denominazioni delle Viste generate dall'Archivio di interesse per le varie annualità e disponibili in SIM. Per ciascuna Vista si possono consultare i tracciati. La denominazione delle Viste è esplicativa rispetto alla presenza di dati personali, al riferimento temporale, inserito come suffisso, e al tipo di unità in esse contenute (Individui, Unità economiche, Luoghi), come mostrato nel Prospetto 2.6.

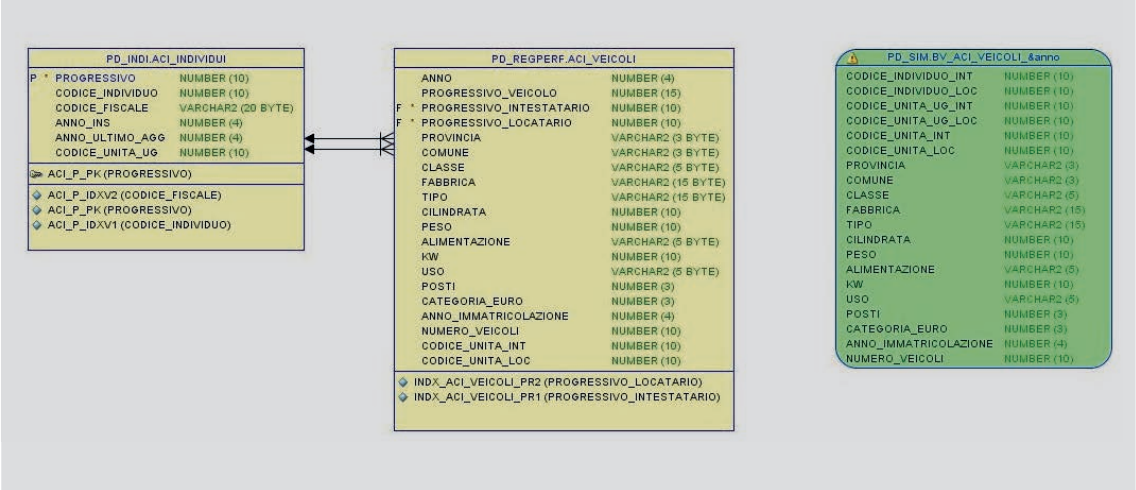
Prospetto 2.6 - Denominazioni delle Viste di SIM in relazione al Tipo di dati

Tipologia dei dati in relazione alle possibilità di accesso	Prefisso del nome della Vista	Descrizione	Unità		
			Unità economiche	Individui	Individui e Unità economiche
Dati pseudonimizzati (solo codici SIM)	BV	Vista che contiene i codici SIM ma non le variabili identificative	✓	✓	✓
	BVDS	Vista che contiene dati sensibili, i codici SIM ma non le variabili identificative	✓	✓	✓
Dati con variabili identificative (senza codici SIM)	SV	Vista che contiene le variabili identificative ma priva dei codici SIM	✓	✓	✓
	SVDS	Vista che contiene dati sensibili, variabili identificative ma priva dei codici SIM	✓	✓	✓
Dati con variabili identificative e pseudonimi (con codici SIM)	V	Vista che contiene il codice SIM e le variabili identificative per le Unità economiche	✓		✓
	VDS	Vista che contiene dati sensibili, i codici SIM e le variabili identificative per le Unità economiche	✓		✓
	IV	Vista che contiene i codici SIM e le variabili identificative		✓	✓
	IVDS	Vista che contiene dati sensibili, i codici SIM e le variabili identificative		✓	✓
-	CUI	Vista contenente il Codice Unico dell'Indirizzo	✓	✓	✓

Per dare visibilità all'organizzazione dei dati nel DB SIM, vengono messi a disposizione anche i metadati relativi ai contenuti delle Tabelle Oracle in cui sono conservati i dati dell'Archivio. Le Tabelle sono partizionate per anno e i dati delle forniture, una volta acquisiti, vengono in esse caricati. Nel caso di variazioni nel tempo dei tracciati record di input (nuove variabili acquisite, variabili non più acquisite), la Tabella comprende l'insieme di tutti i campi via via acquisiti. Sono presenti anche campi creati dagli amministratori del sistema per organizzare i dati o per codificare le unità amministrative (Codici SIM) e la codifica Istat delle variabili territoriali quando necessaria (codici Istat di Provincia e Comune).

Una visione d'insieme delle Tabelle Oracle dell'Archivio e delle Viste da esse generate è fornita dal grafico dello schema dei dati prodotto nel DB Oracle: in giallo sono riportati i contenuti delle Tabelle di cui si compone l'Archivio, con le relative chiavi e connessioni; in verde sono riportati i campi delle Viste per gli utenti. Nella seguente figura l'esempio dell'Archivio Parco veicolare dell'ACI mostra le due tabelle del DB sulla sinistra e la Vista messa a disposizione degli utenti dell'Archivio sulla destra.

Figura 2.2 - Report METADATI->Chiarezza/Interpretabilità->Variabili amministrative-> Strutture dei dati in SIM per l'Archivio Parco veicolare dell'ACI



Per completare la documentazione delle variabili dell'Archivio viene proposto uno specifico report sulle Classificazioni amministrative contenute nella fornitura. Queste sono le tabelle di decodifica delle eventuali variabili categoriche presenti. È possibile visualizzare, per ciascuna classificazione, il nome, l'elenco delle modalità che la compongono (codice e descrizione) e, per ogni modalità, è presente l'anno in cui questa è stata inserita per la prima volta.

Sempre in merito alle classificazioni amministrative, vengono riportate le eventuali connessioni tra classificazioni amministrative (classificazione innestate) o le trascodifiche disponibili nel database SIM – relative, ad esempio, agli stati esteri – e che supportano la fase di pseudonimizzazione.

In uno specifico report *METADATI->Altra documentazione* è accessibile per gli utenti la documentazione, eventualmente disponibile, che possa essere utile al corretto utilizzo dei dati degli archivi. Si tratta di documentazione tecnica in formati vari (pdf, xls, doc, etc.) inviata dall'Ente titolare della Fonte, modulistica amministrativa, testi di Convenzioni che ratificano lo scambio dei dati, tracciati record concordati per l'invio dei dati, e così via. La documentazione è storicizzata rispetto al riferimento temporale dei dati. Questo report è particolarmente utile per gli Archivi che non sono trattati in SIM e la cui documentazione, quindi, non beneficia dei metadati di processo derivanti dalle operazioni ETL di caricamento dei dati nel DB di SIM. Nella Figura 2.3 un quadro d'insieme della navigazione tra i vari report presenti nell'Iperdimensione dei METADATI.

Figura 2.3 - I report implementati nell'Iperdimensione METADATI

2.3.2 Le altre Dimensioni dell'Iperdimensione METADATI

La seconda dimensione della qualità contenuta nell'Iperdimensione METADATI è la *Comparabilità*. Con il termine Comparabilità si intende il *mapping* dei concetti amministrativi e dei concetti statistici. Tale comparazione è necessaria perché, a volte, le definizioni delle variabili e delle unità amministrative spesso non coincidono con le corrispondenti definizioni statistiche. In questa dimensione si vuole dare conto del grado di comparabilità che esiste tra queste: se siano uguali, se si possa operare una trasformazione o se i concetti siano invece incomparabili. Questa dimensione non è attualmente implementata nella QRCA.

Rispetto alla dimensione della *Stabilità temporale dei concetti amministrativi*, uno specifico focus considera i cambiamenti che possono determinare impatti sugli utilizzi statistici dei dati, a causa di modifiche normative o per variazioni tecniche delle fonti. Al fine di gestire in modo ottimale i cambiamenti, oltre a sensibilizzare i titolari dei dati amministrativi a comunicare in anticipo eventuali modifiche previste[3], è opportuno, monitorare i metadati acquisiti. Attualmente nella QRCA non è presente uno specifico report e sono mostrate le informazioni relative alle unità e alle variabili disponibili nell'ultima fornitura ricevuta. Una funzione di monitoraggio è presente nell'Iperdimensione dei DATI, dimensione dei *Controlli tecnici*, in cui è possibile visualizzare alcuni indicatori in serie storica e verificare la presenza di valori anomali (cfr. § 2.4).

3 L'art. 2 comma 2 lett. c) del Decreto del Presidente della Repubblica 7 settembre 2010 n. 166 "Regolamento recante il riordino dell'Istituto Nazionale di Statistica" ha affidato all'Istat il compito di "definire i metodi e i formati da utilizzare da parte delle pubbliche amministrazioni per lo scambio e l'utilizzo in via telematica dell'informazione statistica e finanziaria nonché" di "coordinare modificazioni, integrazioni e nuove impostazioni della modulistica e dei sistemi informativi utilizzati dalle pubbliche amministrazioni per raccogliere informazioni utilizzate o da utilizzare per fini statistici".

L'ultima dimensione della qualità nell'ambito dell'Iperdimensione dei METADATI concerne i possibili trattamenti sui dati effettuati dall'Ente: in questo caso è bene acquisire tutte le informazioni disponibili al fine di poter correttamente utilizzare i dati per i fini statistici e documentarne la qualità. Attualmente non si hanno informazioni per questi aspetti e nella QRCA non è presente uno specifico report.

2.4 L'Iperdimensione DATI

L'Iperdimensione dei DATI contiene i concetti di qualità specifici dei dati una volta che questi vengono acquisiti. Per questo motivo il primo report proposto è lo Stato delle forniture in cui sono presenti le caratteristiche strutturali dei dataset acquisiti e lo stato del monitoraggio delle acquisizioni. Le successive dimensioni valutano gli altri aspetti presenti nel framework adottato. Le Dimensioni in complesso sono le seguenti:

- Stato delle forniture;
- Aspetti temporali;
- Controlli tecnici;
- Integrabilità/Integrazione;
- Accuratezza e coerenza interna;
- Completezza.

2.4.1 Stato delle forniture

Il report delle Stato delle forniture permette di entrare nel merito dei concetti connessi all'identificazione dei dati da acquisire e al loro stato. La Fornitura dei dati amministrativi, come già riportato nel paragrafo 1.1.1, è costituita da uno o più dataset amministrativi ricevuti dall'Istituto nazionale di statistica in seguito ad una specifica richiesta. In questo report, per ciascun Archivio, sono elencati i dataset amministrativi che compongono ciascuna Fornitura. In particolare le Forniture sono caratterizzate dalle informazioni/misure di seguito riportate.

Prospetto 2.7 - Indicatori, informazioni/misure del report DATI - Stato delle forniture

INDICATORE	Informazione/Misura
Descrizione dei dataset (input dei processi statistici)	Nome della fornitura
	Periodicità della fornitura
	Numero d'ordine del dataset nel caso di forniture multiple nell'anno
	Riferimento puntuale dei dati
	Anno di riferimento
	Data minima per la ricezione della fornitura
	Data massima per la ricezione della fornitura
	Modalità di acquisizione
	Trattamento previsto
	Stato del monitoraggio
	Data dello stato del monitoraggio
	Note

Oltre al nome, ciascuna Fornitura è caratterizzata dalla periodicità con cui viene inviata all'Istat: Annuale, Semestrale, Trimestrale, Mensile, Altro (per le forniture occasionali o con periodicità irregolare). La Periodicità viene definita nella fase di Identificazione dei fabbisogni statistici, in seguito ad un confronto tra l'Istat e l'Ente titolare dell'Archivio al fine di conciliare le necessità statistiche con le disponibilità dell'Ente. Conseguentemente, il singolo dataset è identificato da un progressivo; ad esempio, ad una fornitura trimestrale corrisponderanno i dataset 1,2,3,4 per i vari trimestri e così via per le altre periodicità. Ovviamente l'altro elemento caratterizzante è l'anno di riferimento dei dati. Conclude l'identificazione temporale, se necessario, l'informazione del riferimento puntuale del giorno, "dati al gg/mm", completato dall'Anno di riferimento. Se non è valorizzato si assume implicitamente che i dati della fornitura si riferiscano alla fine di ciascun periodo temporale (ad esempio per le Forniture annuali il giorno di riferimento è implicitamente il 31/12), questo dato è utile per il calcolo della Tempestività che vedremo in seguito. Nel caso in cui Periodicità assume la modalità "Altro", il riferimento del giorno è necessario alla corretta identificazione del dataset, e indica, esplicitamente, l'ultimo giorno del periodo di riferimento dei dati.

Un elemento utile alla caratterizzazione dei dataset sarebbe anche la data di estrazione dalla Fonte amministrativa per verificare la puntualità della registrazione degli eventi. La data di estrazione permetterebbe di identificare la Fornitura di dati provvisori" dalla "Fornitura di dati definitivi" che hanno lo stesso riferimento temporale.

Purtroppo, tale informazione non è disponibile, si considera come sua *proxy* la data di acquisizione. In questi casi, per individuare agevolmente le tipologie di forniture, i termini "dati provvisori" e " dati definitivi" sono inseriti nel nome della fornitura. Sempre nel caso di due forniture aventi lo stesso riferimento temporale ma estratte in due momenti diversi dalla Fonte, si utilizza il termine "aggiornamento dei dati" piuttosto che "dati definitivi" per indicare il caso in cui i due dataset siano complementari nel periodo.

Oltre ai riferimenti temporali, la fornitura è caratterizzata dagli accordi con l'Ente titolare in merito al processo di acquisizione, ovvero l'intervallo temporale in cui è previsto che la fornitura venga inviata. La data minima e la data massima dell'intervallo sono concordate tra le esigenze dei processi di produzione e le disponibilità dell'Ente.

Per quanto riguarda l'informazione sulla Modalità di acquisizione, le modalità sicure secondo gli standard dell'Istat, sono elencate nel Prospetto 2.8.

Prospetto 2.8 - Modalità sicure utilizzate per l'acquisizione dei dati amministrativi

MODALITA' DI ACQUISIZIONE	Descrizione
Portale	I dati sono inviati utilizzando il portale ARCAM
Canale SFTP	I dati sono acquisiti attraverso il canale protetto SSH File Transfer Protocol
Web	I dati sono acquisiti tramite un'applicazione web specifica
Pec e altro	I dati sono inviati tramite PEC o con modalità protette diverse da quelle indicate (solo per gli archivi che non contengono dati personali)
Open data	L'Ente titolare mette a disposizione dati di tipo aperto attraverso servizi web a cui l'Istat accede

Il report dello *Stato delle forniture*, oltre a caratterizzare i dataset, ha lo scopo di informare sullo Stato del monitoraggio in cui essi si trovano: l'informazione dello Stato, affiancato alla Data dello Stato, indica il punto in cui si trova la fornitura nell'ambito del ciclo di vita all'interno del processo di gestione centralizzata dei dati amministrativi. I possibili stati che attraversa la fornitura dipendono dal trattamento, previsto per l'Archivio di riferimento. Per gli Archivi non trattati in SIM gli stati possibili delle forniture sono i seguenti.

Prospetto 2.9 - Stati del monitoraggio delle forniture per gli Archivi non trattati da SIM

STATO	Descrizione
In fase di richiesta all'Ente	La fornitura è stata inserita nella Programmazione annuale delle acquisizioni ed è stata avviata all'iter di richiesta all'Ente
Richiesta all'Ente, in attesa di acquisizione	La fornitura è stata richiesta ufficialmente all'Ente ed è in attesa di acquisizione secondo le modalità e i tempi concordati
Trasmessa dall'Ente, in fase di acquisizione	La fornitura è arrivata in Istat e i dati sono in attesa di transitare dall'area temporanea all'area dedicata alla memorizzazione
Acquisita	La fornitura è stata acquisita

Per le Forniture che vengono sottoposte a trattamento in SIM, gli stati del monitoraggio possibili sono riportati nel Prospetto 2.10; alcuni sono comuni a tutte le forniture, altri si differenziano in relazione alla specifica Tipologia di trattamento prevista. L'elenco dei trattamenti è descritto nel. Paragrafo 2.2.1.

Prospetto 2.10 - Stati del monitoraggio delle forniture per gli Archivi trattati da SIM

STATO	Descrizione
In fase di richiesta all'Ente	La fornitura è stata inserita nella Programmazione annuale delle acquisizioni ed è stata avviata all'iter di richiesta all'Ente
Richiesta all'Ente, in attesa di acquisizione	La fornitura è stata richiesta ufficialmente all'Ente ed è in attesa di acquisizione secondo le modalità e i tempi concordati
Trasmessa dall'Ente, in fase di acquisizione	La fornitura è arrivata in Istat e i dati sono in attesa di transitare dall'area temporanea all'area dedicata alla memorizzazione
Acquisita, da trattare in SIM	La fornitura è stata acquisita ed è in attesa di trattamento nel SIM (Sistema Integrato di Microdati)
Acquisita, trattata in SIM e in attesa di assegnazione dei codici SIM	La fornitura è stata trattata nel SIM (Sistema Integrato di Microdati), ovvero i metadati sono stati inseriti o aggiornati e i dati sono stati caricati nelle tabelle del DB, sono state effettuate le codifiche di provincia e comune, laddove necessario; è in attesa di ricevere i codici SIM
Acquisita, trattata in SIM con prima assegnazione del codice SIM Unità giuridica, in attesa di assegnazione dei codici SIM	La fornitura è stata trattata nel SIM (Sistema Integrato di Microdati), ovvero i metadati sono stati inseriti o aggiornati e i dati sono stati caricati nelle tabelle del DB, sono state effettuate le codifiche di provincia e comune, laddove necessario; è stato apposto preliminarmente il Codice Unità Giuridica ed è in attesa di ricevere i codici SIM
Acquisita, trattata in SIM con assegnazione del codice SIM Unità Economica, in attesa di assegnazione del codice SIM Individuo	La fornitura è stata trattata nel SIM (Sistema Integrato di Microdati), ovvero i metadati sono stati inseriti o aggiornati e i dati sono stati caricati nelle tabelle del DB, sono state effettuate le codifiche di provincia e comune, laddove necessario; è stato apposto il Codice Unità Economica ed è in attesa di ricevere il Codice Individuo
Acquisita, trattata in SIM con prima assegnazione del Codice SIM Unità giuridica e con assegnazione codice SIM Individuo, in attesa di assegnazione del Codice SIM Unità economica	La fornitura è stata è stata trattata nel SIM (Sistema Integrato di Microdati), ovvero i metadati sono stati inseriti o aggiornati e i dati sono stati caricati nelle tabelle del DB, sono state effettuate le codifiche di provincia e comune, laddove necessario; è stato apposto il Codice Individuo e, preliminarmente, il Codice Unità Giuridica ed è in attesa di ricevere il codice Unità economica
Disponibile	La fornitura è disponibile all'utente
Disponibile con codici SIM	La fornitura è disponibile all'utente con i Codici SIM

A titolo di esempio, si presenta nel Prospetto 2.11 una parte del report *DATI->Stato delle Forniture* per l'Archivio Dati su energia elettrica e gas naturale dell'Ente ARERA, non trattato da SIM. Dal prospetto si evince che le Forniture Annuali riferite agli anni 2017-2019 sono Disponibili per gli utenti, mentre la Fornitura del 2020 è in Fase di richiesta all'Ente.

Prospetto 2.11 - Monitoraggio dell'Archivio Dati su energia elettrica e gas naturale nel report della QRCA DATI - Stato delle Forniture (informazione al 02/07/2021)

NOME FORNITURA	Periodicità	Anno di Riferimento	Trattamento previsto	Stato Monitoraggio	Data Stato Monitoraggio
Dati su energia elettrica e gas naturale	Annuale	2020	Nessun trattamento in SIM	In fase di richiesta all'ente	
Dati su energia elettrica e gas naturale	Annuale	2019	Nessun trattamento in SIM	Disponibile	11/23/2020
Dati su energia elettrica e gas naturale	Annuale	2018	Nessun trattamento in SIM	Disponibile	11/25/2019
Dati su energia elettrica e gas naturale	Annuale	2017	Nessun trattamento in SIM	Disponibile	11/28/2018

Mentre il monitoraggio di un Archivio trattato in SIM, come l'Archivio dei Lavoratori interinali dell'INAIL, presenterà nel report *DATI->Stato delle forniture*, la situazione riportata nel Prospetto 2.12.

Prospetto 2.12 - Monitoraggio dell'Archivio Lavoratori interinali nel report della QRCA DATI - Stato delle Forniture (informazione al 02/07/2021)

Nome Fornitura composto	Periodicità	Anno di Riferimento	Trattamento previsto	Stato Monitoraggio	Data Stato Monitoraggio
Lavoratori interinali	Annuale	2020	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	In fase di richiesta all'ente	
Lavoratori interinali	Annuale	2019	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	Disponibile con codici SIM	12/21/2020
Lavoratori interinali	Annuale	2018	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	Disponibile con codici SIM	12/10/2019
Lavoratori interinali	Annuale	2017	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	Disponibile con codici SIM	1/15/2019
Lavoratori interinali	Annuale	2016	Trattamento in SIM con assegnazione del Codice SIM a livello di individuo e di unità economica	Disponibile con codici SIM	1/11/2018

Conclude il report il campo delle Note inserite dall'Ente titolare in fase di invio della fornitura dei dati o dal sistema QRCA per descrivere particolari caratteristiche dei dati utili all'utente.

2.4.2 Aspetti temporali

La seconda dimensione della qualità comprende gli Aspetti temporali che determinano la qualità dei dati: la Tempestività e la Puntualità. Rispetto ai tradizionali indicatori associati alla diffusione dell'output delle Rilevazioni statistiche, per i dati amministrativi il significato si applica alla Puntualità dell'Ente che fornisce i dati e alla tempestività rispetto all'input dei processi.

Prospetto 2.13 - Indicatori, informazioni/misure del report DATI - Aspetti temporali

INDICATORE	Informazione/Misura
Puntualità	Confronto tra la data di acquisizione della fornitura e l'intervallo temporale concordato per l'invio dei dati
Tempestività dell'Ente	Intervallo temporale tra la data di acquisizione dei dati e l'ultimo evento registrato nei dati della fornitura
Tempestività complessiva	Intervallo temporale tra la data di disponibilità dei dati pretrattati centralmente e l'ultimo evento registrato nei dati della fornitura

La QRCA calcola, in base ai dati disponibili, l'indicatore di Puntualità per ciascuna fornitura acquisita, ovvero il rispetto delle date di consegna concordate con l'Ente titolare in fase di richiesta dei dati.

La misura della puntualità è data dal confronto tra la data effettiva di acquisizione dei dati e l'intervallo temporale concordato per la consegna dei dati [Data minima; Data massima]. In particolare:

- Se Data effettiva di acquisizione ∈ [Data minima - Data massima] =>
  - Stato della puntualità = 'Puntuale'
  - Misura di puntualità = 0
- Se Data effettiva di acquisizione < Data minima =>
  - Stato della puntualità = 'Anticipo'
  - Misura di puntualità = [*Data effettiva di acquisizione - Data minima*] (*numero di giorni*)
- Se Data effettiva di acquisizione > Data massima =>
  - Stato della puntualità = 'Ritardo'
  - Misura di puntualità = [*Data effettiva di acquisizione - Data massima*] (*numero di giorni*)

L'indicatore di tempestività è presente in due misure: Tempestività dell'Ente e Tempestività complessiva. La tempestività dell'Ente considera i valori della tempestività di consegna della fornitura da parte dell'Ente, ovvero quanto tempo intercorre tra la consegna e il periodo a cui si riferiscono i dati contenuti nella fornitura.

Misura della tempestività dell'Ente = [Data effettiva di acquisizione dei dati in Istat - Data finale del periodo di riferimento dei dati] (numero di giorni)

dove la Data finale del periodo di riferimento dei dati considera la data dell'ultimo evento registrato nei dati della fornitura. In assenza di informazioni specifiche è determinata come ultimo giorno del periodo (31 dicembre dell'anno di riferimento se la fornitura è annuale, 30 giugno dell'anno di riferimento se la fornitura è semestrale e così via).

La tempestività complessiva calcola la differenza tra la data di disponibilità della fornitura agli utenti Istat, a seguito dell'eventuale trattamento dei dati in SIM, e la data finale del periodo di riferimento dei dati.

Se i dati non sono trattati in SIM, la tempestività dell'Ente coincide con la tempestività complessiva.

Misura della tempestività complessiva = [Data di disponibilità dei dati per i processi di produzione - Data finale del periodo di riferimento dei dati] (numero di giorni)

2.4.3 Controlli tecnici

L'Iperdimensione dei DATI comprende la valutazione dei dati acquisiti. I *Technical check* o Controlli tecnici hanno lo scopo di verificare la conformità dei dati ricevuti, rispetto ai dati attesi in seguito alla richiesta ufficiale da parte dell'Istat; l'esito di tali controlli è una prima forma di validazione della fornitura dei dati; in caso si riscontrino dei problemi, il più celermente possibile, si procede a ricontattare l'Ente per chiedere informazioni o, quando necessario, il rinvio i dati.

I report contenuti in questa dimensione mostrano i risultati dei controlli effettuati sui dati e gli indicatori vengono presentati in serie storica per fornire anche indicazioni sul monitoraggio della qualità dei dati e identificare in modo tempestivo la possibile presenza di anomalie.

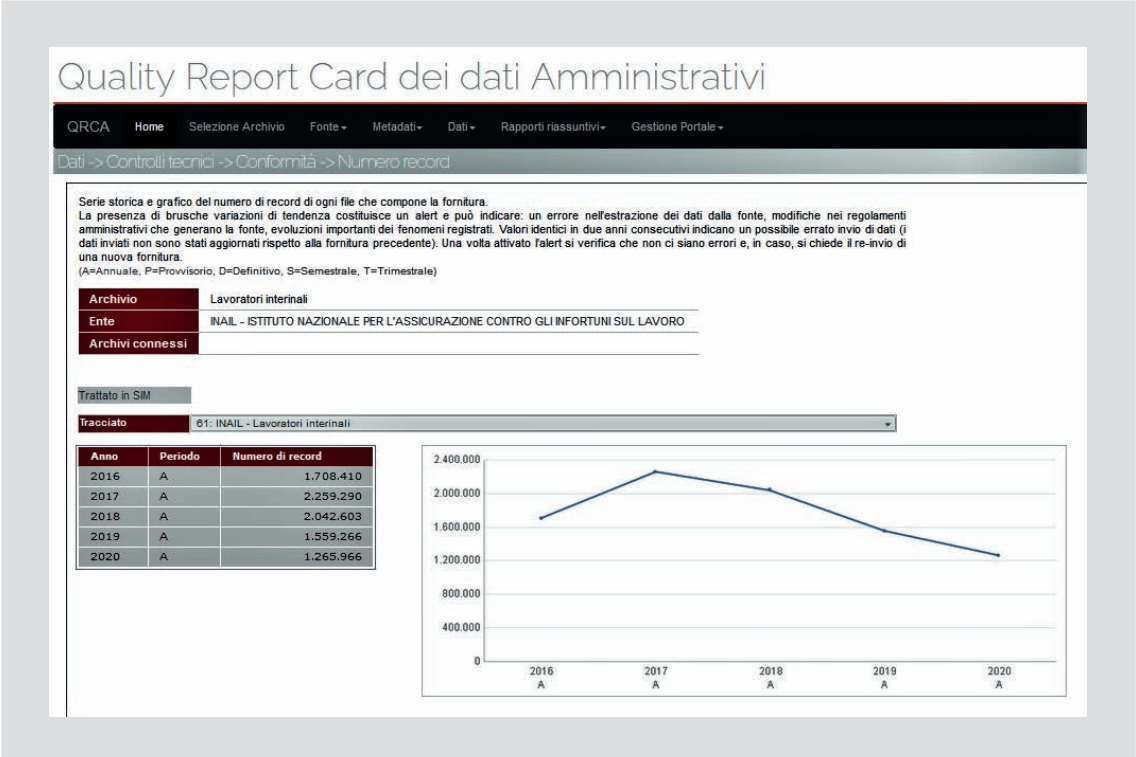
Prospetto 2.14 - Indicatori, informazioni/misure del report DATI - Controlli tecnici

INDICATORE	Informazione/Misura
Conformità	Numero di record per ciascun file che compone la fornitura, valori assoluti in serie storica
Conformità	Percentuale dei campi valorizzati sul totale dei record valorizzati in serie storica per ogni variabile fornita
Conformità	Frequenze percentuali delle modalità in serie storica per ogni variabile categorica
Conformità	Modalità con decodifiche mancanti e corrispondente numero di record in serie storica per ogni variabile categorica

Il primo controllo consiste nel contare il numero di record per ciascun file che compone la fornitura e compararlo con la serie storica, ciò permette di verificare se il nuovo conteggio rispetta il trend in relazione alle forniture precedenti.

L'indicatore viene restituito in forma numerica e in forma grafica per permettere un veloce controllo visivo (vedi Figura 2.4).

Figura 2.4 - Serie storica del numero di record per il file 61 - Lavoratori interinali della Fornitura dell'Archivio dei Lavoratori interinali dell'INAIL

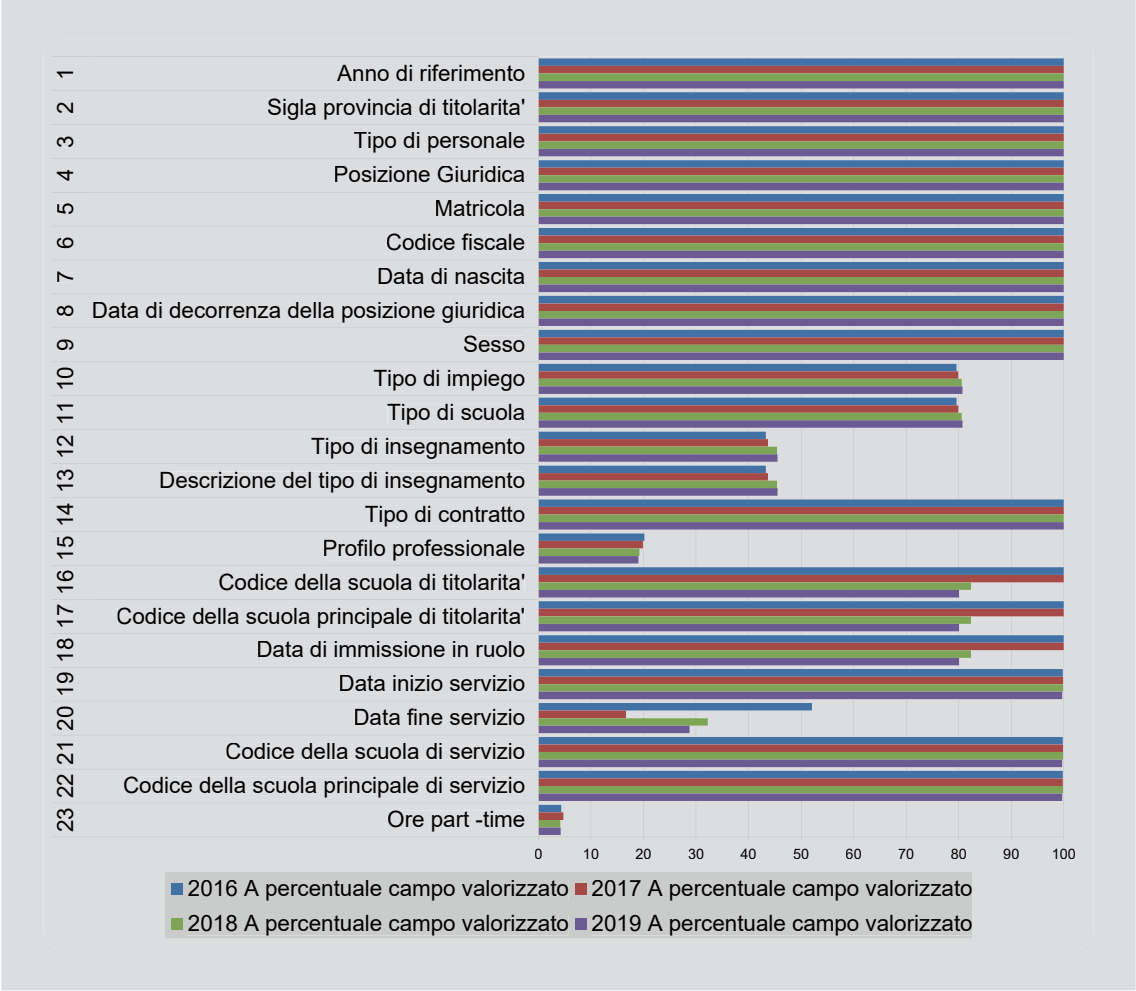


La presenza di brusche variazioni di tendenza costituisce un *alert* e può indicare: un errore nell'estrazione dei dati dalla Fonte, evoluzioni importanti dei fenomeni registrati, modifiche nei regolamenti amministrativi che generano la Fonte. Valori identici in due anni consecutivi indicano un possibile errato invio di dati (i dati inviati non sono stati aggiornati

rispetto alla fornitura precedente). Una volta attivato l'*alert,* si verifica che non ci siano errori e, in caso, si chiede il rinvio di una nuova fornitura. In generale, sarebbe opportuno che l'Ente titolare della Fonte avvisasse l'Istat in caso di variazioni note a priori; purtroppo ciò non sempre accade e una funzione di monitoraggio come quella dei Controlli tecnici può risultare molto utile.

Il secondo dei Controlli tecnici verifica la completezza della fornitura: per ogni file della fornitura e per ogni variabile si visualizza la serie storica della Percentuale delle valorizzazioni sul totale dei record. Per facilitare la lettura, si rende disponibile anche graficamente, il confronto con i valori relativi alle Forniture precedenti. Nella seguente Figura 2.5, si riporta l'esempio dei controlli effettuati sull'Archivio dell'Anagrafe del personale scuole statali e istituti comprensivi per il file Archivio del personale delle scuole.

Figura 2.5 - Percentuale dei campi valorizzati delle variabili della Fornitura dell'Archivio dell'Anagrafe del personale scuole statali e istituti comprensivi



Anche in questo caso, variazioni inattese in serie storica delle valorizzazioni possono indicare possibili errori nella fornitura la cui origine potrà essere esplorata presso l'Ente.

Il terzo *check* effettuato per valutare la conformità dei dati acquisiti controlla le variabili categoriche che l'Ente ha fornito corredate da una tabella di classificazione che trascodifica i codici delle modalità: per queste variabili la QRCA elabora le frequenze percentuali delle

modalità. Il corrispondente report, per ogni file e per ogni variabile categorica, presenta in tabella, le percentuali in serie storica, calcolate come il numero di valorizzazioni di ciascuna modalità/il numero totale delle valorizzazioni della variabile.

Anche in questo caso, brusche variazioni delle composizioni percentuali possono segnalare un possibile errore, la cui origine può essere tempestivamente indagata. Nei casi di errore, si richiede all'Ente di rinviare la fornitura.

Sempre per le variabili categoriche, nell'ultimo controllo si analizza la presenza di decodifiche mancanti nella tabella di trascodifica dei codici delle modalità. Decodifiche mancanti possono essere dovute alla presenza di valori fuori range, oppure alla mancanza di aggiornamento delle tabelle di trascodifica. Il report mostra, per ogni file della fornitura e per ogni variabile categorica, la serie storica delle modalità non presenti nella tabelle delle trascodifiche e il numero di record associati.

2.4.4 Integrabilità/Integrazione dei dati

I dati amministrativi esprimono tutta la loro potenzialità di uso a fini statistici con l'integrazione. Quando i dati amministrativi contengono microdati riferiti alle unità target della statistica, è possibile integrare i dati in senso longitudinale oppure dati appartenenti ad Archivi diversi. Ad esempio, si possono integrare le iscrizioni nei vari anni scolastici, o i dati sui livelli di istruzione con i dati sui redditi, i dati sul lavoro con i dati dei componenti della famiglia e così via.

Al fine di poter combinare i dati amministrativi e adempiere alla normativa vigente sul trattamento dei dati personali, l'Istat sta completando l'adozione di una serie di procedure di protezione *by design* e *by default*. Innanzitutto i dati amministrativi acquisiti dall'Istat vengono classificati in due tipologie, a seconda che contengano o meno Dati Personali.

Per gli archivi che contengono Dati Personali viene applicata la procedura di pseudonimizzazione, definitiva dal Regolamento UE 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016, art.4: come "il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile".

La procedura di pseudonimizzazione viene effettuata, come visto nel paragrafo 2.1.1, nell'ambito dei trattamenti SIM. Il processo comprende i seguenti passi:

1. Separazione degli identificativi presenti nel dataset dalle variabili tematiche. Le variabili che permettono l'identificazione delle unità statistiche di base vengono conservate separatamente e l'accesso ai dati è rigidamente regolato.
2. Riconoscimento delle unità statistiche di base (Individuo e/o Impresa) attraverso procedure di *record linkage* di tipo deterministico e apposizione del codice pseudonimo, per gli Individui e per le Unità economiche.
   Per ogni Archivio, in base alle variabili identificative disponibili e alla loro qualità, si utilizza una specifica procedura di *record linkage* che abbina le unità presenti nel dataset con la lista delle unità già riconosciute negli anni nel SIM, se una Rappresentazione dell'unità[4] viene abbinata ad una già presente, si associa lo stesso codice pseudonimo;

4 La Rappresentazione dell'unità presente in un record del dataset è l'insieme dei valori assunti dalle variabili identificative fornite.

se la Rappresentazione non si abbina, l'unità è un nuovo ingresso nel sistema, viene assegnato un nuovo codice (per maggiori approfondimenti sulla procedura si veda Runci, Di Bella, Galiè, 2016).

La fase di messa a disposizione dei singoli dataset da parte della struttura centralizzata di gestione dei dati amministrativi prevede uno specifico governo degli accessi. I dati personali vengono rilasciati in forma pseudonima (le variabili identificative sono sostituite dallo pseudonimo) e gli pseudonimi resi disponibili permettono, dove necessario per le finalità statistiche, di integrare i dataset. Laddove, per i processi di produzione, sia indispensabile accedere anche agli Identificativi, si definisce l'accesso al set minimo necessario previa la dichiarazione nella delibera di accesso ai dati personali del trattamento da effettuare. In futuro, l'introduzione in SIM di codici pseudonimi plurimi, secondo un approccio gerarchico per domini specifici di integrazione, permetterà di migliorare ulteriormente la gestione degli accessi ai dati seguendo i rilievi espressi dal Garante per la protezione dei dati personali (Garante sulla protezione dei dati personali, 2020).

Questa operazione permette di minimizzare il rischio di identificabilità degli interessati quando non strettamente necessario per i fini statistici.

Nell'Iperdimensione della qualità dei DATI, vengono proposti alcuni indicatori che descrivono le potenzialità di integrazione dei dati (Linkabilità delle unità) e i risultati del processo di pseudonimizzazione dei dati personali effettuato nel SIM per gli archivi che contengono microdati a livello di individuo. La valutazione della qualità della procedura di pseudonimizzazione considera l'attività di identificazione delle unità statistiche attraverso le procedure di record linkage applicate. Nel Prospetto 2.15 sono elencate le Misure presenti nella QRCA per gli Archivi trattati con apposizione del codice SIM a livello di Individuo.

Prospetto 2.15 - Indicatori, informazioni/misure del report DATI - Integrabilità/Integrazione

INDICATORE	Informazione/Misura
Linkailità delle unità	Numero ed elenco delle variabili disponibili e utilizzate come chiavi di linkage per l'identificazione e la pseudonimizzazione delle unità
Linkabilità delle unità	Percentuale di valori mancanti sul totale dei record per ogni variabile di linkage disponibile nella fornitura
Accuratezza delle unità/monitoraggio del processo di integrazione	Serie storica della distribuzione percentuale di record per Tipo di abbinamento in SIM al momento del record linkage (ultimi due anni)
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con individui che non si sono abbinati in SIM al momento del record linkage e a cui è stato associato un nuovo Codice Individuo
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con individui che si sono abbinati in SIM al momento del record linkage
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con individui abbinati in SIM al momento del record linkage e persistenti nell'Archivio
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con individui abbinati in SIM al momento del record linkage e non persistenti nell'Archivio (nuovi abbinati)

L'indicatore di linkabilità delle unità Individuo è presente attraverso due misure: la prima indica la disponibilità per l'Archivio delle variabili identificative utilizzate come chiavi di linkage per l'assegnazione del Codice Individuo nel SIM con lo scopo di pseudonimizzare i dati personali quando presenti.

L'insieme delle variabili di linkage utilizzate dalle procedure sono: cognome; nome; sesso; data di nascita; luogo di nascita; cittadinanza; codice fiscale; indirizzo di residenza o di domicilio fiscale.

Il secondo indicatore di qualità riporta la percentuale di valori mancanti sul totale dei record per ciascuna variabile di linkage disponibile, calcolata sui dati dell'ultima fornitura disponibile.

La qualità del processo di integrazione, è in genere, proporzionale rispetto al numero di variabili di linkage disponibili e alla loro qualità.

Nel Report del monitoraggio del processo di integrazione sono mostrate alcune misure di accuratezza delle unità calcolate utilizzando i metadati del processo di pseudonimizzazione.

Il processo di identificazione delle unità (Runci *et al.*, 2016) prevede che al primo passo si ricerchi per similitudine l'intera Rappresentazione tra tutte le Rappresentazioni già presenti nel SIM, ad esempio, la Rappresentazione dell'Individuo [RSSMRA70S30A062Z Mario Rossi Roma Affile 30/11/1970]). Le Rappresentazioni che si abbinano vengono etichettate con lo STEP=1. Per i record residui si procede con il passo 2 che solitamente lascia fuori il codice fiscale, e così via per le altre variabili. Le Rappresentazioni che non trovano un abbinamento in SIM ricevono un nuovo Codice Individuo. Per le forniture successive alla prima, un passo preliminare riconosce gli Individui Persistenti nell'Archivio, ovvero presenti nel dataset con la stessa Rappresentazione e attribuisce loro lo stesso codice Individuo assegnato nella fornitura precedente solo se il primo abbinamento è stato effettuato al passo 1 (STEP=1).

Le misure proposte sono presentate in serie storiche, corredate da grafici, al fine di monitorare questa fase ed evidenziare possibili anomalie o errori nei dati di input o nel processo di identificazione. La prima misura riporta la distribuzione percentuale di record per Tipo di abbinamento in SIM al momento dell'esecuzione della procedura di record linkage. Le modalità sono:

- individui persistenti;
- record abbinati allo Step 1 - abbinamento per similitudine di tutte le variabili identificative presenti nell'Archivio;
- record abbinati allo Step 2;
- record abbinati allo Step 3;
- … ;
- non abbinati a cui viene assegnato un nuovo Codice Individuo.

Il significato della variabile Step n per n=2,3,.., varia da Archivio ad Archivio e dipende da quanti e quali chiavi di linkage sono disponibili e dalla loro qualità. L'interpretazione della misura si basa sull'evidenza che quando l'Archivio mette a disposizione molte variabili identificative, l'abbinamento al primo passo garantisce una buona qualità di linkage (ad esempio quando le rappresentazioni abbinate hanno stesso codice fiscale, cognome, nome, sesso, data di nascita, luogo di nascita hanno necessariamente una buona qualità di linkage).

Oltre alla tabelle dei dati, si forniscono una serie di grafici:

- il confronto dei valori delle distribuzione degli ultimi due anni;
- la serie storica della percentuale di record con Individui che non si sono abbinati in SIM e a cui è stato associato un nuovo Codice Individuo;
- la serie storica della percentuale di record con Individui che si sono abbinati in SIM;
- la serie storica della percentuale di record con Individui abbinati in SIM al momento comprensivi dei persistenti;
- la serie storica della percentuale di record con Individui abbinati in SIM ma non persistenti nell'Archivio (nuovi abbinati).

2.4.5 Le altre Dimensioni dell'Iperdimensione DATI

L'Accuratezza ha l'obiettivo di misurare l'inconsistenza dei dati per le unità, per le relazioni, per le variabili e loro combinazione. Nella dimensione dell'Integrabilità/Integrazione sono presentate alcune misure dell'accuratezza per le unità. Altre misure, al momento, non sembrano documentabili secondo la strategia della QRCA, ovvero utilizzando metadati di processo già esistenti poiché sarebbe necessario attingere ai sistemi dei processi di produzione che non sono standardizzati e raccolti in un'unica piattaforma. Un'ipotesi da esplorare in futuro, potrebbe essere di considerare dei processi statistici di riferimento per mettere a disposizione degli altri processi che utilizzano lo stesso Archivio le prime fasi di controllo dei dati ed evitare duplicazioni di attività.

Nella dimensione della Completezza dei dati si misurano due aspetti: la completezza rispetto alle unità, ovvero gli indicatori di copertura; la completezza rispetto alle variabili, ovvero la presenza di valori mancanti. L'indicatore di completezza delle variabili è derivabile dai Controlli tecnici, Percentuale dei campi valorizzati, mentre gli indicatori di copertura degli Archivi non sono implementati nella QRCA poiché attualmente non è possibile generarli periodicamente per tutti i dataset in modo automatico. Come per gli indicatori di Accuratezza, in futuro, sarebbe utile condividere le misure di controllo calcolate su specifici processi, come ad esempio i processi di produzione dei Registri statistici dell'Istat e mettere a disposizione degli altri processi i risultati della valutazione.

In Appendice sono riportate le misure non ancora implementate nell'Iperdimensione dei DATI.

3. LA STRATEGIA DI PRODUZIONE DELLA QRCA

3.1 Il riuso dei metadati di processo

Il principio fondante della QRCA è l'efficienza del processo di produzione e l'utilità delle informazioni prodotte. In Istat, già tentativi di documentazione dei dati amministrativi precedenti alla QRCA erano stati abbandonati per le difficoltà connesse alle operazioni di aggiornamento. Le informazioni da considerare, infatti, sono numerose e possiedono un certo grado di dinamicità, lo scenario di base ha le seguenti caratteristiche:

- gli indicatori di qualità devono essere prodotti per circa 190 Archivi amministrativi e per un numero che va da 400 a 500 forniture annue che si cumulano nel sistema arrivando a circa 1.900 forniture a fine anno 2021 (Tavola 1.1);
- le caratteristiche dei dataset amministrativi hanno un'elevata variabilità in termini di formato, contenuto, struttura dei dati;
- i dataset amministrativi sono spesso molto grandi in termini di *byte*;
- quando presenti dati personali, occorre operare nel rispetto della normativa in termini di riservatezza e trattamento dei dati.

Sulla base di tale scenario e dei fabbisogni informativi dell'Istat, una volta definito il framework teorico (cfr. § 1.3), si è effettuato uno studio di fattibilità per verificare le ipotesi possibili di implementazione di un sistema di documentazione.

Il risultato di tale valutazione ha indotto a seguire il principio chiave che prevede "*Reuse metadata where possible for statistical integration as well as efficiency reasons*" definito nell'ambito del Gruppo di supporto per la modernizzazione della Statistica Ufficiale (Unece, 2009).

Quindi, nel percorso che ha portato alla produzione della QRCA, la prima attività ha riguardato l'analisi di fattibilità della strategia: ovvero verificare se i metadati degli IT Tool di gestione dei dati amministrativi, in particolare Arcam e SIM, fossero riutilizzabili ai fini della documentazione, se fossero interoperabili e sufficienti per lo scopo.

Come già detto, Arcam è il sistema di acquisizione e, attraverso il suo DB, gestisce le informazioni per popolare le due interfacce: tra il fornitore dei dati amministrativi e l'Istat e tra il sistema e gli amministratori (back office). SIM ha l'obiettivo di pretrattare i dati amministrativi che contengono dati personali e gestisce i processi di ETL e di pseudonimizzazione.

La quantità di informazione disponibile su Arcam è subito sembrata interessante, anche se i livelli di standardizzazione necessari al loro riuso ai fini della pubblicazione sono risultati inadeguati: i metadati pur assolvendo i compiti della funzione di acquisizione dovevano seguire un passaggio di standardizzazione concettuale e formale. I metadati di processo di SIM hanno mostrato un elevatissimo livello di riuso dal momento in cui le procedure sono altamente automatizzate ed esiste il sottosistema di gestione dei metadati di processo denominato SISME, solo alcune attività sono state ulteriormente standardizzate ai fini del riuso. L'inserimento di poche specifiche Tabelle e Viste create per seguire i trattamenti ne hanno decretato l'usabilità.

L'analisi di fattibilità ha riguardato anche l'interoperabilità dei due sistemi. L'entità comune ai due sistemi è l'Archivio amministrativo e i due sistemi contengono le loro liste di riferimento, diverse tra loro perché nate con finalità distinte ma anche perché basate su concetti diversi tra loro dal punto di vista operativo, pur se riconducibili alla definizione del Glossario della QRCA. Si tratta, quindi, della struttura concettuale di un dataset, con tutte le proprietà caratterizzanti, che definisce il sottoinsieme delle informazioni della Fonte amministrativa da utilizzare per le finalità statistiche dell'Istat.

Operativamente, il concetto di Archivio in Arcam rispetta la convenienza dell'Ente per l'invio dei dati. Occorre considerare che, nel rispetto del principio di minimizzazione del fastidio statistico o *response burden* anche per i fornitori di dati amministrativi, gli Enti sono posti nelle condizioni di inviare i dati nel modo a loro più consono evitando di richiedere loro specifiche elaborazioni che necessiterebbero di risorse aggiuntive e potrebbero, inoltre, essere fonte di errore.

Il concetto di Archivio di SIM è legato all'individuazione delle Entità che rappresentano la popolazione di interesse e dalle loro relazioni, per poter organizzare la struttura dei dati corrispondente.

Questa diversità ha dato luogo ad un processo di verifica dell'interoperabilità, risolta con successo attraverso una semplice tabella di raccordo che riconduce una lista all'altra e risolve le eventuali relazioni (N:M), anch'esse documentate nei report della QRCA. Ad esempio, l'Archivio fornito dall'INPS delle Certificazioni telematiche di malattia che in Arcam viene considerato come unico, si divide in SIM nei due archivi: Certificazioni telematiche di malattia dipendenti pubblici, Certificazioni telematiche di malattia dipendenti privati: nella tabella di raccordo si registra la relazione.

Ovviamente la tabella di raccordo deve essere aggiornata in occasione dell'acquisizione di nuovi archivi.

Oltre ai metadati di processo, si sono aggiunti alcuni metadati del Dizionario del DB Oracle e specifici macrodati di supporto al calcolo degli indicatori definiti appositamente da una procedura che si aggiorna quando un nuovo dataset viene caricato nel SIM.

Ulteriori apporti alla QRCA sono forniti dal DB del Programma Statistico Nazionale (Psn) e dal sistema SIQual, il Sistema Informativo sulla Qualità delle rilevazioni ed elaborazioni condotte dall'Istat[1] per il calcolo degli indicatori di Rilevanza. Per quanto riguarda il Psn l'informazione di interesse riguarda l'associazione tra lavoro Psn dell'Istat e l'Archivio amministrativo utilizzato che, nella compilazione delle Schede del Psn da parte dei referenti statistici, deve essere indicato quando utilizzato come input del processo. Per il Sistema SIQual si sfrutta l'associazione tra lavoro Istat del Psn e la Normativa europea che il lavoro statistico deve adempiere che il sistema documenta. Nel paragrafo 3.4 sono descritte le relazioni con i due sistemi che permettono il riuso dell'informazione nella QRCA.

Il punto di forza del sistema di produzione della QRCA è che l'aggiornamento sistematico delle sole tabelle di raccordo attiva automaticamente il simultaneo aggiornamento degli oggetti documentati.

In sintesi, le attività svolte per giungere alla costruzione operativa della QRCA sono state le seguenti:

a) Adozione del framework della qualità dei dati amministrativi;
b) Definizione degli indicatori e delle misure di qualità;
c) Analisi dei processi di gestione dei dati amministrativi, dei metadati disponibili nei sistemi IT e del flusso dei dati;

1 Il sistema di navigazione è accessibile dal sito esterno dell'Istat al seguente link: http://siqual.istat.it.

d) Studio di fattibilità per l'interoperabilità e il riuso dei metadati e classificazione delle misure in:
   - Implementabili nel breve periodo con i metadati già esistenti
   - Implementabili nel medio periodo con i metadati esistenti ma ancora non accessibili
   - Implementabili nel lungo periodo con informazioni da acquisire
e) Implementazione dell'interoperabilità;
f) Scelta degli strumenti tecnologici;
g) Definizione delle specifiche tecniche, costruzione e test del primo prototipo della QRCA.

Grazie a questa strategia la QRCA può seguire il ciclo di vita dei dati amministrativi: nella fase di descrizione e acquisizione dei dati, attraverso il sistema Arcam, le informazione del DB del Psn e del Sistema SIQual; nella fase di trattamento e rilascio, attingendo ai metadati di processo di SIM.

Dal punto di vista informatico, la QRCA si basa su una piattaforma di *Visual Analytics* di *Business Intelligence* - MicroStrategy, strumento adottato dall'Istat per realizzare alcuni prodotti di visualizzazione dei dati. Il sito web e le sue funzionalità sono gestiti da un'applicazione Java generalizzata. Il flusso delle informazioni viene supportato da un DB Oracle che comprende le Viste delle Tabelle di interesse dei Sistemi di origine e altre Tabelle proprie che supportano il software di produzione per il calcolo degli indicatori (Calabria *et al.*, 2018).

3.2 Le relazioni tra la QRCA e il Sistema Arcam

La lista degli archivi presente nella QRCA proviene dal sistema Arcam, che contiene le informazioni a livello di Archivio e di Fornitura in relazione al processo di acquisizione dei dati amministrativi. Attraverso delle specifiche Viste sul DB di Arcam, la QRCA seleziona gli Archivi e le Forniture da documentare insieme alle caratteristiche necessarie alla loro corretta identificazione e descrizione.

L'aggiornamento dei metadati di identificazione degli Archivi e delle Forniture è garantito in Arcam dalle attività di acquisizione che prevedono la definizione del Programma annuale di acquisizione delle forniture (§1.2 Fase di Specificazione dei fabbisogni) e che determina l'insieme e le caratteristiche delle Forniture da acquisire nell'anno. Quando necessario, nel corso dell'anno, la programmazione viene integrata da ulteriori Forniture. Queste attività vengono svolte nell'ambito della Direzione della raccolta dati da uno specifico team composto dai Referenti dell'acquisizione i quali gestiscono i metadati attraverso l'apposita interfaccia di *back-office* di Arcam.

Le informazioni di monitoraggio delle acquisizioni, data di acquisizione e stato del monitoraggio, vengono aggiornate automaticamente dal sistema ogni volta che si realizza un cambio di stato delle forniture: dallo stato iniziale "In fase di richiesta all'Ente, allo stato di "Richiesta all'Ente, in fase di acquisizione", fino a quando avviene la trasmissione dei dati da parte dell'Ente e il sistema aggiorna allo stato transitorio di "Trasmessa dall'Ente, in attesa di acquisizione" e infine allo stato di "Acquisita" con corrispondente valorizzazione delle date dei vari stati attraversati. Se la fornitura deve essere trattata in SIM la QRCA provvede a modificare lo stato finale in "Acquisita, da trattare in SIM (cfr. Prospetto 2.10).

Occorre considerare però che, in alcuni casi, la trasmissione dei dati da parte degli Enti all'Istat può avvenire attraverso canali diversi dal Portale Arcam (Prospetto 2.8). Ad esempio, buona parte degli Archivi di Inps e Agenzia delle Entrate sono trasmessi mediante il canale SFTP, un protocollo di rete che prevede il trasferimento dei dati in modalità sicura.

Se l'invio dei dati della Fornitura non avviene mediante Portale, il referente dell'Ente inserisce sul Portale di Arcam, le informazioni relative alla trasmissione della Fornitura (numero dei file, natura dei file, presenza di tracciato o di altra documentazione) e il Referente Istat dell'acquisizione aggiorna manualmente lo stato della Fornitura in "Acquisita" e inserisce la data di acquisizione. Questo fa sì che, indipendentemente dal canale di trasmissione utilizzato, in Arcam siano presenti le informazioni riguardanti il monitoraggio dell'acquisizione di tutti gli archivi documentati in QRCA.

Poiché il sistema Arcam è stato realizzato prima dell'implementazione della QRCA, con la finalità di ottenere un sistema per la trasmissione sicura dei dati e per il monitoraggio delle acquisizioni è stato necessario, come detto, un processo di standardizzazione dei metadati già presenti nel sistema per garantirne la corretta riusabilità rispetto alle loro funzioni di origine. Al fine di armonizzare i metadati di Arcam è stata effettuata una prima analisi della casistiche e sono state definite le strategie di intervento. L'attività di armonizzazione ha coinvolto tutti i Referenti dell'acquisizione. Inoltre, sono state definite delle linee guida destinate ai referenti di Arcam in modo tale da assicurare la corretta gestione dei metadati nel corso del tempo.

La standardizzazione dei metadati in Arcam ha riguardato principalmente i seguenti aspetti:

- armonizzazione delle denominazione di Archivi, Forniture e Enti titolari,
- integrazione e coerenza delle variabili relative ai riferimenti temporali delle forniture
- inserimento dei concetti di Fornitura per la quale sono previsti più invii distintamente tra dati provvisori/dati definitivi e dati parziali/integrazioni
- corretta individuazione delle forniture trasmesse dagli Enti che risultano incomplete o errate.

I report che utilizzano, in parte, i metadati di processo di Arcam sono: *FONTE->Informazioni di base, DATI->Stato delle forniture, DATI->Aspetti temporali-> Puntualità e tempestività, Rapporti riassuntivi ->Monitoraggio*.

3.3 Le relazioni tra la QRCA e il sistema SIM

Come già evidenziato nei precedenti paragrafi, il ciclo di vita dei dati amministrativi acquisiti dall'Istat presso il Servizio centralizzato prevede che alcuni Archivi vengano rilasciati direttamente agli utenti interni titolari dei processi di produzione, mentre altri, generalmente quelli che contengono microdati relativi alle unità statistiche di base Individui, Unità Economiche e Luoghi, che – dal punto di vista normativo si può tradurre nella presenza di Dati Personali – vengano pretrattati e sottoposti a procedure di protezione.

Il SIM svolge la funzione di pretrattamento dei dati amministrativi, ovvero tutte le operazioni di tipo trasversale comuni ai processi che utilizzano i dati amministrativi con lo scopo di supportarli. Inoltre, permette una rappresentazione omogenea rispetto alla varietà delle diverse sorgenti, rende i dati più aderenti alla logica del Sistema, ne garantisce la persistenza e consente agli utenti di disporre di un'astrazione di alto livello attraverso cui interagire.

In linea generale le sue funzioni comprendono: la gestione dei singoli dataset ovvero l'esecuzione delle procedure di ETL, precedute dall'analisi E/R dei dati; la gestione dei metadati; le procedure di identificazione delle unità target della statistica; la protezione dei dati personali per il rilascio in sicurezza ai processi di produzione statistica dell'Istat.

Le procedure, scritte per gli Archivi al momento della prima acquisizione e manutenute in occasione dei cambiamenti che si possono presentano nel tempo (nuove variabili, nuovi formati, nuove strutture dei dati), sono parametrizzate – laddove possibile – considerando la grande varietà che caratterizza i dataset amministrativi che giungono all'Istat. Oltre alle Tabelle dei Dati, SIM comprende una serie di Tabelle di Metadati utili a svolgere il pretrattamento: sono dei metadati strutturali che consentono il caricamento standardizzato e veloce dei dati nelle Tabelle e dei metadati di processo, ovvero delle procedure e delle funzioni che vengono lanciate nelle varie fasi del pretrattamento.

L'elevato livello di generalizzazione del sistema, come già detto, ha permesso di mettere in atto la strategia della QRCA e il riuso dei metadati garantisce il popolamento di diversi report.

Le procedure di ETL permettono di documentare i contenuti degli archivi trattati: la QRCA, accede a specifiche Viste su SISME e MicroStrategy permette il popolamento dei vari Report dell'Iperdimensione dei METADATI.

I metadati del processo relativi alle operazioni di pseudonimizzazione, come riportato nel paragrafo 2.4.4, popolano i report della Dimensione *DATI->Integrabilità Integrazione*.

Il *Data Dictionary* di Oracle, collezione di viste e sinonimi di sistema che fornisce informazioni dettagliate sulla struttura, sulle attività e sugli oggetti contenuti nel DB, supporta ulteriormente la QRCA nella produzione della documentazione. Ad esempio, l'elenco e il contenuto delle Viste predisposte per il rilascio dei dati è individuato automaticamente attraverso l'uso della Vista ALL_DEPENDENCIES associata alle Tabelle SISME.

Per la funzione di monitoraggio dei trattamenti, a valle dei quali lo stato della fornitura risulta Disponibile, si utilizzano degli specifici metadati di processo di SIM che documentano le varie fasi del processo e le date di sistema associate.

Per quanto riguarda i report dei *DATI->Controlli tecnici*, una specifica procedura della QRCA che attinge ai dataset da caricare, calcola i macrodati necessari per la produzione del report (numero di record, valori missing, frequenze delle variabili categoriche per ciascun dataset amministrativo).

3.4 Le relazioni tra la QRCA e i sistemi Psn e SIQual

Per l'implementazione dell'Indicatore di Rilevanza la QRCA si avvale del riuso delle informazioni provenienti dai due sistemi informativi: il Psn e SIQual.

Come già visto, il Psn è l'atto normativo che regolamenta i processi statistici espletati nell'ambito del Sistan, La definizione di ciascun lavoro statistico del Psn avviene mediante la compilazione di un'apposita scheda in cui vengono riportate le principali informazioni del processo statistico, tra le altre: descrizione dei principali fenomeni osservati, tipologia di dati utilizzati, natura dei dati personali trattati, caratteristiche del rilascio dei dati). Nell'ambito della tipologia dei dati trattati vengono elencati gli Archivi amministrativi che, in caso, servono all'espletamento del processo statistico. La QRCA accede alle informazioni delle schede dei lavori Psn di titolarità dell'Istat acquisendo l'elenco degli Archivi dichiarati per ciascun lavoro. In particolare utilizza le seguenti informazioni:

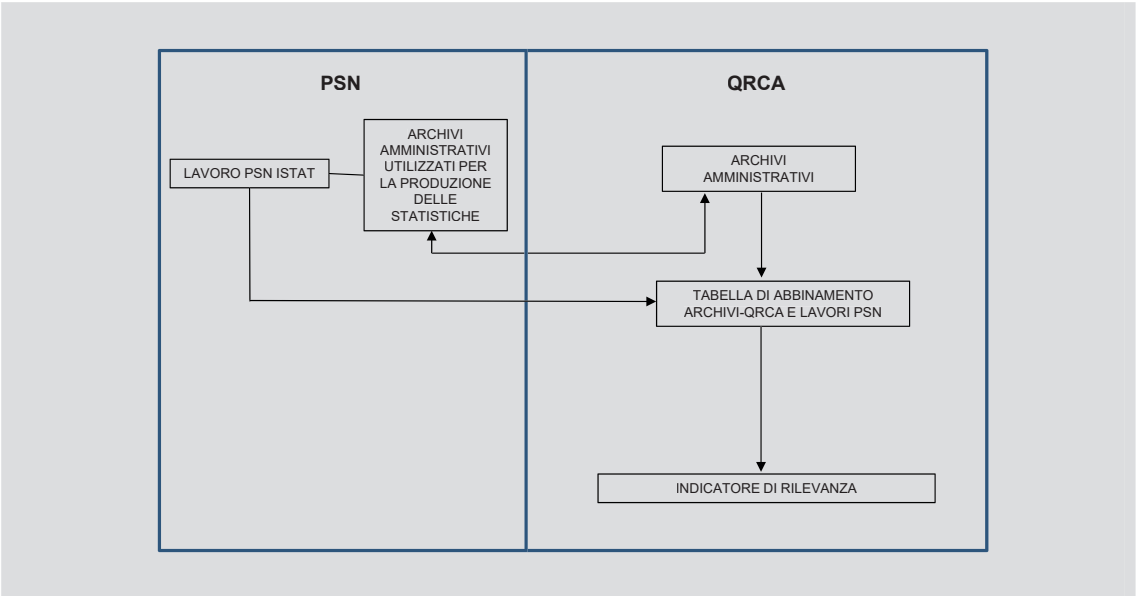
- codice e denominazione del lavoro Psn;
- codice e denominazione degli archivi amministrativi utilizzati;
- struttura incaricata allo svolgimento del lavoro Psn (dipartimento, direzione e servizio dell'Istat).

Al fine di poter riutilizzare questi metadati nell'Indicatore di Rilevanza degli Archivi amministrativi, la QRCA elabora una tabella di trascodifica tra l'elenco degli Archivi del Psn e l'elenco degli Archivi di Arcam.

Il processo di abbinamento è risultato abbastanza complesso a causa della forte disomogeneità delle denominazioni degli archivi Psn e del loro disallineamento con quanto riportato in Arcam. Tale situazione è legata sia al fatto che il sistema Psn è aperto a tutti i soggetti del Sistan sia al fatto che la sua realizzazione è antecedente alla centralizzazione della Raccolta Dati in Istat, pertanto la compilazione delle schede del Psn era gestita in maniera indipendente dalle strutture di produzione che avevano a proprio carico tutte le fasi del processo statistico.

Nella Figura 3.1 è schematizzato il processo di riuso dei dati del Psn.

Figura 3.1 - Relazione tra la QRCA e il Psn



L'informazione derivata dal Psn viene arricchita da un ulteriore contributo del sistema SIQual. In esso la documentazione delle rilevazioni ed elaborazioni condotte dall'Istat comprende anche l'informazione relativa all'eventuale Normativa comunitaria che regola la produzione delle statistiche. Il vincolo normativo, ovviamente, fornisce al processo un maggiore livello di importanza e può arricchire la rilevanza di un archivio utilizzato come input del processo stesso. Operativamente, poiché in SIQual i processi sono individuati tramite un codice e ciascun codice è associato a uno o più lavori Istat del Psn, se lo svolgimento dell'indagine è vincolato all'adempimento della normativa comunitaria, i lavori Psn associati ereditano l'attributo della presenza della Normativa comunitaria. La QRCA, utilizzando la chiave del codice del lavoro Psn già acquisita dal DB del Psn (Figura 3.1), è così in grado di associare tale attributo all'Archivio pubblicando le misure della Rilevanza dell'Archivio il cui uso è connesso con l'adempimento della Normativa europea.

3.5 L'interoperabilità dei sistemi per la caratterizzazione dei dati di input

La chiave di abbinamento tra i sistemi Arcam, SIM e Psn è il codice Archivio, raccordato dalle rispettive tabelle di transcodifica, mentre il collegamento con SIQual è garantito dal passaggio attraverso il codice del lavoro Psn.

Psn e SIQual forniscono informazioni a livello di Archivio permettendo di descrivere la parte iniziale del ciclo di vita del dato amministrativo, laddove i lavori statistici del Psn esprimono il fabbisogno di dati amministrativi attraverso la compilazione della Scheda del Psn (§1.2 Fase di Specificazione dei fabbisogni). SIQual rafforza il fabbisogno attraverso la necessità di adempimento delle Normative europee e la QRCA ne ricava gli Indicatori di Rilevanza.

Al fine di poter seguire più in dettaglio il ciclo, occorre entrare al livello della Fornitura; in questo caso le informazioni derivano dall'interoperabilità tra Arcam e SIM e, oltre al codice Archivio, la QRCA utilizza anche i riferimenti temporali e i concetti armonizzati di Fornitura e File della Fornitura attraverso il riconoscimento dei file di SIM appartenenti alla stessa Fornitura di Arcam e dei file appartenenti a Forniture diverse.

Dall'insieme dei metadati dei due sistemi così riconnessi si deriva l'informazione più importante dal punto di vista operativo, ovvero il monitoraggio del flusso dei dati. Gli utenti Istat della QRCA, attraverso il report *DATI->Stato delle Forniture*, possono verificare in tempo reale il percorso dei dati: dalle prime fasi dell'acquisizione al trattamento fino allo stato di disponibilità (cfr. Prospetti 2.9 e 2.10).

Il seguente Prospetto 3.1 mostra una schematizzazione degli oggetti e delle relazioni coinvolti. Per ogni oggetto/relazione sono riportati il sistema in cui gli oggetti e le relazioni hanno origine e le attività svolte dalla QRCA per garantire l'interoperabilità tra i sistemi e mantenere il filo che segue il ciclo di vita dei dati amministrativi. In quest'ottica, il Sistema della QRCA può essere considerato un sistema Multifonte poichè la documentazione prodotta deriva da più fonti: Arcam, SIM, Psn, SIQual.

Prospetto 3.1 - Oggetti e sistemi informativi dell'Istat in cui l'oggetto è originato

OGGETTO/RELAZIONE	Sistema	Azioni della QRCA
Archivio amministrativo	ARCAM	Seleziona gli Archivi da documentare
Lavori statistici	PSN	Seleziona i lavori statistici dell'Istat che utilizzano dati amministrativi
Fornitura	ARCAM	Seleziona le Forniture da documentare
File di dati	SIM	Individua i file di SIM relativi agli Archivi e alla Forniture da documentare
Normativa europea	SIQUAL	Seleziona la Normativa associata ai lavori PSN che utilizzano dati amministrativi
Archivio – Lavoro statistico	QRCA	Determina e aggiorna l'abbinamento attraverso una tabella di trascodifica
Archivio – Normativa europea	QRCA	La QRCA associa gli Archivi alla Normativa europea
File – Fornitura	QRCA	La QRCA associa i file di SIM con le Forniture degli Archivi da documentare e mantiene aggiornata l'associazione

3.6 I controlli e la manutenzione del sistema

Affinché sia garantito il corretto aggiornamento della QRCA, occorre manutenere le tabelle di connessione tra i sistemi Arcam e SIM in occasione di specifici eventi che de-

terminano l'inserimento di un nuovo elemento o il cambiamento delle caratteristiche di un elemento già presente nel sistema. I principali eventi sono di seguito elencati:

- acquisizione di un nuovo Archivio che sarà trattato in SIM;
- acquisizione di una nuova fornitura fuori programmazione che andrà in SIM;
- inserimento in SIM di un Archivio già acquisito e presente in QRCA;
- cambiamento del tipo di trattamento di un Archivio già trattato in SIM;
- riprogrammazione dell'invio di una fornitura previsto per l'anno corrente.

Seguendo una logica prestabilita, codificata in apposite Linee Guida, viene intrapreso un determinato insieme di azioni al presentarsi di ciascun evento di cambiamento. Gli amministratori della QRCA aggiornano le tabelle di connessione tra i sistemi e verificano che le operazioni di gestione effettuate in Arcam e in SIM siano tempestive ed allineate.

La collaborazione con gli amministratori di Arcam e gli amministratori di SIM è fondamentale per mantenere la standardizzazione dei metadati di processo che vengono pubblicati o che permettono il corretto funzionamento degli algoritmi di calcolo delle misure e degli indicatori: al fine di poter riusare i metadati di processo, la gestione delle operazioni di acquisizione e di trattamento devono essere svolte con un minimo di attenzione alla funzione derivata della documentazione nella QRCA. Una buona collaborazione garantisce anche la tempestività dell'aggiornamento dei report della QRCA. Si sottolinea, d'altra parte, che gli operatori dell'acquisizione e del trattamento beneficiano anch'essi del processo di documentazione e ne sono, quindi, ripagati (cfr. § 1.4).

Poiché le azioni sono sequenziali e coinvolgono più soggetti, per avere la sicurezza dell'esaustività e della correttezza di tutti i passaggi, vengono periodicamente eseguite più serie di controlli automatizzati. I controlli permettono di verificare se la procedura di aggiornamento ha funzionato e, in caso di necessità, consentono di agire rapidamente per la risoluzione di eventuali anomalie. Il set di controlli varia a seconda dell'evento di cambiamento e delle relative azioni che vengono intraprese e agiscono sui sistemi di origine o, in ultima istanza, sui report della QRCA.

A titolo di esempio, vengono descritti nel dettaglio le azioni e i controlli relativi al primo evento di cambiamento.

Quando viene stabilita l'acquisizione di un nuovo Archivio che dovrà essere trattato da SIM, gli amministratori di Arcam inseriscono le informazioni relative all'Archivio e alle Forniture ad esso associate. Il Sistema genera automaticamente i codici identificativi dell'Archivio e delle relative Forniture. Contemporaneamente gli amministratori di SIM inseriscono i primi metadati nelle tabelle di SISME mentre gli amministratori della QRCA provvedono ad aggiornare le tabelle di raccordo. A questo punto la documentazione del nuovo Archivio e le corrispondenti forniture saranno automaticamente visibili nella QRCA e si avvia la documentazione di monitoraggio dell'acquisizione e del trattamento. Una volta acquisita la prima fornitura di dati, la definizione delle procedure di ETL completano il popolamento dei metadati di SISME e di conseguenza i relativi Report della QRCA.

I controlli che vengono eseguiti hanno lo scopo di accertare il corretto inserimento delle informazioni a livello di Archivio e di Fornitura in Arcam, in SIM e nelle tabelle di raccordo. Al momento dell'acquisizione, si verifica che le informazioni integrate provenienti da Arcam e SIM popolino in modo corretto e coerente il report del monitoraggio delle forniture.

CONCLUSIONI E SVILUPPI FUTURI

Dal 2018 – anno di nascita della QRCA – il sistema dimostra nel tempo una buona tenuta e un buon livello di flessibilità rispetto alla capacità di adeguamento alle modifiche dei processi e alle esigenze della produzione statistica.

Anche l'inserimento delle nuove misure nella versione 2.0, volta ad ampliare le informazioni necessarie a descrivere più in dettaglio gli archivi, le forniture e gli accessi, hanno richiesto piccole modifiche alle procedure di produzione del sistema. Si osserva, inoltre, una buona scalabilità in termini quantitativi, considerando che il numero degli oggetti da documentare cresce nel tempo: nuovi archivi vengono acquisiti e le forniture periodiche si cumulano nei report, mentre gli archivi non più acquisiti continuano ad essere documentati.

In questo periodo, l'Istat sta progettando un sistema di governance e di controllo generale del sistema di acquisizione e pretrattamento dei dati amministrativi con lo scopo di mettere in atto delle soluzioni tecnologiche e organizzative per realizzare la piena compliance al Regolamento generale sulla protezione dei dati personali (Regolamento (Ue) 2016/679 - aggiornato alle rettifiche pubblicate sulla Gazzetta Ufficiale dell'Unione europea 127 del 23 maggio 2018.). Questa attività nasce per dare risposta ad una serie di rilievi del Garante sulla protezione dei dati personali[1] che sottolineano la necessità di gestire la protezione dei dati *by design* e *by default* e, specificatamente, adottare idonee tecniche di pseudonimizzazione in SIM per garantire l'effettività dei principi di minimizzazione e di limitazione della conservazione.

La QRCA, ovviamente, è coinvolta sia in maniera diretta che indiretta in questa fase progettuale. Il suo coinvolgimento è diretto per la sua funzione di documentazione del ciclo di vita dei dati amministrativi, inoltre, il lavoro svolto per standardizzare la terminologia, la descrizione del flusso dei dati, ha contribuito a focalizzare le fasi di acquisizione, monitoraggio, trattamento e valutazione della qualità coinvolti nella nuova progettazione. Il coinvolgimento indiretto è dovuto alle modalità di alimentazione dei suoi report: laddove cambieranno le procedure e i processi di trattamento occorre garantire la continuità della funzione di documentazione, innovando se necessario gli strumenti IT. L'occasione è indubbiamente stimolante per la possibilità di arricchire ulteriormente la QRCA e dare massima trasparenza ai fabbisogni, agli usi e alla qualità dei dati e dei processi, requisito fondamentale per la protezione dei dati.

Un ulteriore elemento di stimolo per gli sviluppi futuri della QRCA deriva dai recenti progressi nella costruzione in Istat del Sistema dei Registri statistici. I Registri statistici sono alimentanti per la gran parte da dati di Fonte amministrativa e il loro output statistico in forma di microdati fornisce la possibilità di acquisire degli interessanti feedback per la valutazione della qualità dell'input. Un esempio per tutti è la possibilità di calcolare gli indicatori di copertura (sottocopertura e sovracopertura) di una fornitura amministrativa rispetto alla popolazione target statistica. Gli indicatori di copertura, calcolati periodicamente per l'anno di riferimento T a valle del rilascio dei dati del relativo Registro statistico, possono essere documentati nella QRCA in serie storica.

1 Per dare conto del processo in corso si cita l'ultimo parere disponibile, Garante sulla protezione dei dati personali (2020).

Un'ultima osservazione riguarda le potenzialità di inserimento di ulteriori indicatori nella QRCA: delle 106 misure riportate in Appendice, 57 sono definite teoricamente e potrebbero essere implementate in futuro. Per queste misure è possibile considerare alcune categorie definite in relazione alla tipologia di sostenibilità dell'implementazione.

Una prima categoria comprende quelle informazioni che, pur facilmente reperibili, richiederebbero un aggiornamento manuale; ad esempio nell'Iperdimensione FONTE, l'Identificazione della Fonte potrebbe essere descritta attraverso gli eventuali Regolamenti amministrativi o specifiche gestionali che generano la Fonte o attraverso la pubblicazione in QRCA dei Moduli utilizzati per l'acquisizione/registrazione dei dati. Un'altra informazione appartenente alla stessa categoria è la documentazione delle relazioni e dei feedback con l'Ente titolare della Fonte, come ad esempio Protocolli d'Intesa, Accordi, Gruppi di lavoro definiti tra l'Istat e l'Ente. Queste misure richiedono un presidio costante per verificarne l'aggiornamento con un notevole aumento delle risorse necessarie e una maggiore possibilità di introdurre errori nel sistema.

Un altro insieme di misure è caratterizzato dal fatto che la loro implementazione generalizzata sarebbe troppo onerosa e richiederebbe competenze afferenti ai processi di produzione degli output statistici esterne alla Direzione preposta alla raccolta dati che ha in carico la QRCA, ad esempio le misure di *Accuratezza* nell'Iperdimensione DATI. Ciò non toglie che, per alcuni Archivi diffusamente utilizzati dall'Istat, alcune informazioni, come ad esempio gli esiti di controlli di coerenza sui dati, determinati nell'ambito di un processo statistico di riferimento, non possano essere messe a disposizione di tutti gli utenti Istat nella QRCA: in tal caso la fonte dell'informazione sarebbe il singolo processo. Questa strategia si potrebbe applicare in futuro con lo scopo di evitare duplicazioni di attività nell'Istituto. In generale, si può affermare che il confine della QRCA potrebbe essere determinato dall'insieme delle misure di qualità dell'input dei processi che mantengano una loro generalità rispetto agli usi e che possano essere considerate utili a tutti i processi che usano l'Archivio oggetto di valutazione.

Accanto a questa categoria di misure, si possono annoverare gli indicatori di *Comparabilità* dell'Iperdimensione METADATI. Per questi sarebbe interessante un confronto con il catalogo delle variabili target di output, disponibile nel Sistema Unico dei Metadati (SUM) dell'Istat (Signore *et al.*, 2015), per determinare le corrispondenze tra le variabili amministrative e le variabili statistiche e tentare una misurazione della distanza dal punto di vista concettuale.

L'idea futura di standardizzazione dei processi potrebbe portare alla documentazione del ciclo complessivo dei processi che utilizzano dati amministrativi passando da un sistema ad un altro, ad esempio connettendo maggiormente la QRCA con il sistema SIQual. Queste strade saranno esplorate in futuro, condizionatamente alle esigenze dell'Istituto.

APPENDICE - INDICATORI, INFORMAZIONI E MISURE DELLA QRCA

Prospetto 1 - Indicatori, informazioni/misure per la Dimensione FONTE

INDICATORE	Informazione/Misura	Da implementare
Identificazione della Fonte	Denominazione dell'Archivio (contiene il nome della Fonte)	
Identificazione della Fonte	Denominazione dell'Ente titolare della Fonte amministrativa	
Identificazione della Fonte	Regolamenti amministrativi o specifiche gestionali che generano la Fonte amministrativa	x
Identificazione della Fonte	Moduli/applicativi utilizzati per l'acquisizione/registrazione dei dati	x
Riservatezza e protezione dei dati	Presenza di dati personali (Reg. UE 2016/678) (SI/NO)	
Riservatezza e protezione dei dati	Presenza di dati rientranti in particolari categorie (ex sensibili) (SI/NO)	
Riservatezza e protezione dei dati	Presenza di dati relativi a condanne penali e reati (ex giudiziari) (SI/NO)	
Lunghezza della serie dei dati	Serie storica dei dati disponibili per l'Archivio (anni disponibili)	
Tipologia di trattamento	Tipologia di trattamento a cui è sottoposto l'Archivio	
Lunghezza della serie dei dati	Serie storica disponibile nel SIM (anni)	
Tipologia di trattamento	Codifica degli indirizzi in RSBL (SI/NO)	
Lunghezza della serie dei dati	Anni in RSBL	
Gestione dell'acquisizione	Referente Istat per l'acquisizione	
Relazioni e feedback con l'Ente titolare della fonte	Protocolli d'intesa, Accordi, Gruppi di lavoro tra l'Istat e l'Ente	x
Relazioni e feedback con l'Ente titolare della fonte	Altre relazioni tra l'Istat e l'Ente	x
Relazioni e feedback con l'Ente titolare della fonte	Costi previsti per l'acquisizione dei dati (SI/NO)	x
Relazioni e feedback con l'Ente titolare della fonte	Informazioni da parte dell'Ente su modifiche previste nella Fonte che possono impattare sulla fornitura dei dati (tracciato record, definizioni di variabili, unità, popolazione target amministrativa, puntualità, tempestività) (SI/NO)	x
Relazioni e feedback con l'Ente titolare della fonte	Grado di collaborazione dell'Ente	x

Prospetto 1 segue **- Indicatori, informazioni/misure per la Dimensione FONTE**

INDICATORE	Informazione/Misura	Da implementare
Rilevanza	Numero dei lavori Istat del Psn che utilizzano i dati dell'Archivio per anno di aggiornamento del Psn (serie storica dal 2017)	
Rilevanza	Elenco e codice dei lavori Istat del Psn che utilizzano i dati dell'Archivio per anno (serie storica dal 2017)	
Rilevanza	Referente statistico Istat (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Scopi di uso statistico Istat (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Numero di normative UE il cui adempimento dipende dall'uso dell'Archivio a fini statistici per lavoro Psn per anno (serie storica dal 2017)	
Rilevanza e usi statistici	Elenco della Normativa UE il cui adempimento dipende dall'uso dell'Archivio a fini statistici per lavoro Psn per anno (serie storica dal 2017)	
Rilevanza	Numero di variabili prodotte con i dati dell'Archivio e non rilevate con l'indagine (tasso di sostituzione) (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Numero di unità non contattate grazie all'uso dei dati amministrativi della Fonte (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Percentuale di riduzione delle risorse e dei tempi grazie all'uso dei dati amministrativi dell'Archivio (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Soddisfazione del fabbisogno informativo in termini di variabili (possibile azione di feedback presso il fornitore per possibili modifiche) (per ogni lavoro Istat che utilizza l'Archivio) (1-piena soddisfazione; 2-migliorabile)	x
Rilevanza	Soddisfazione del fabbisogno informativo in termini di classificazioni (possibile azione di feedback presso il fornitore per possibili modifiche) (per ogni lavoro Istat che utilizza l'Archivio) (1-piena soddisfazione; 2-migliorabile)	x
Rilevanza	Soddisfazione del fabbisogno informativo in termini di unità (per ogni lavoro Istat che utilizza l'Archivio) (1-piena soddisfazione; 2-migliorabile)	x
Rilevanza	Soddisfazione del fabbisogno informativo in termini di tempestività (possibile azione di feedback presso il fornitore per possibili modifiche) (1-piena soddisfazione; 2-migliorabile) (per ogni lavoro Istat che utilizza l'Archivio)	x
Rilevanza	Elenco e codice Psn dei lavori del titolare che utilizzano la Fonte	x

Prospetto 2 - Indicatori, informazioni/misure del report METADATI

INDICATORE	Informazione/Misura	Da implementare
Descrizione dei dati (unità)	Unità e popolazioni presenti nell'Archivio	
Descrizione dei dati (unità)	Definizione degli oggetti (eventi ed unità) e delle popolazioni utilizzabili a fini statistici - indicare l'origine della definizione (da metadati avuti dall'Ente titolare della Fonte, da normativa,...)	x
Descrizione dei dati (variabili)	Elenco delle variabili dell'Archivio fornite dall'Ente	
Descrizione dei dati (variabili)	Definizione delle variabili dell'Archivio - indicare l'origine della definizione (da metadati avuti dal fornitore, da normativa,...)	x
Descrizione dei dati (variabili)	Elenco strutturato delle Viste di SIM per i rilasci interni e delle variabili in esse contenute	
Descrizione dei dati (variabili)	Elenco delle Tabelle Oracle che compongono l'Archivio e delle variabili in esse contenute	
Descrizione dei dati (variabili)	Grafico delle strutture dei dati (Tabelle e chiavi Oracle)	
Descrizione dei dati (variabili)	Classificazioni delle variabili categoriche presenti nell'Archivio	
Descrizione dei dati (variabili)	Descrizione delle modalità delle variabili categoriche presenti nell'Archivio (variabili qualitative/classificazioni)	x
Descrizione dei dati	Altra documentazione disponibile	
Comparabilità	Comparabilità tra gli oggetti amministrativi e le unità statistiche a livello concettuale. Classificazione in [identici, corrispondenti, incomparabili], ovvero comparazione delle definizioni rispetto ad oggetti contenuti in un'altra fonte di riferimento (statistica o amministrativa)	x
Comparabilità	Comparabilità tra le variabili amministrative e le variabili statistiche target a livello concettuale, ovvero comparazione delle definizioni rispetto a variabili contenute in un'altra fonte di riferimento (statistica o amministrativa)	x
Comparabilità	Comparabilità delle classificazioni per le variabili qualitative rispetto a classificazioni presenti in un'altra fonte di riferimento (statistica o amministrativa). [identiche, corrispondenti, incomparabili]	x
Stabilità temporale delle variabili amministrative	Cambiamenti nel tempo dei tracciati record	x
Stabilità temporale delle classificazioni amministrative	Cambiamenti nel tempo delle classificazioni	x
Stabilità temporale delle variabili amministrative identificative	Cambiamenti nella disponibilità delle variabili identificative	x
Stabilità temporale delle popolazioni amministrative	Eventuali cambiamenti nel tempo del campo di osservazione degli oggetti e verifica della disponibilità di documentazione di aggiornamento da parte dell'Ente titolare	x
Informazioni sull'esistenza di piani di controllo o di trattamenti effettuati dall'Ente titolare sui dati	Trattamenti effettuati dal titolare	x

Prospetto 3 - Indicatori, informazioni/misure del report DATI

INDICATORE	Informazione/Misura	Da implementare
Descrizione dei dataset	Nome della fornitura	
Descrizione dei dataset	Periodicità della fornitura	
Descrizione dei dataset	Numero d'ordine del dataset nel caso di forniture multiple nell'anno	
Descrizione dei dataset	Riferimento puntuale dei dati (gg/mm/aaaa)	
Descrizione dei dataset	Anno di riferimento	
Descrizione dei dataset	Data minima per la ricezione della fornitura	
Descrizione dei dataset	Data massima per la ricezione della fornitura	
Descrizione dei dataset	Modalità di acquisizione	
Descrizione dei dataset	Trattamento previsto	
Descrizione dei dataset	Note	
Monitoraggio	Stato del monitoraggio dell'acquisizione e del trattamento	
Monitoraggio	Data dello stato del monitoraggio dell'acquisizione e del trattamento	
Puntualità	Confronto tra la data di acquisizione della fornitura e l'intervallo temporale concordato per l'invio dei dati	
Tempestività dell'Ente	Intervallo temporale tra la data di acquisizione dei dati e l'ultimo evento registrato nei dati della fornitura	
Tempestività complessiva	Intervallo temporale tra la data di disponibilità dei dati pretrattati centralmente e l'ultimo evento registrato nei dati della fornitura	
Lunghezza delle serie storiche	Presenza delle variabili nel tempo nelle varie forniture (serie storica)	x
Dinamicità degli oggetti - Variazioni della popolazione di oggetti nel tempo	Percentuale di oggetti presenti al tempo t ma non al tempo t-1 (nuovi oggetti) rispetto al totale di oggetti al tempo t	x
Dinamicità degli oggetti - Variazioni della popolazione di oggetti nel tempo	Percentuale di oggetti presenti al tempo t-1 ma non al tempo t (vecchi oggetti) rispetto al totale di oggetti al tempo t	x
Dinamicità degli oggetti - Variazioni della popolazione di oggetti nel tempo	Percentuale di oggetti persistenti dal tempo t-1 al tempo t rispetto al totale di oggetti al tempo t	x
Ritardo - Ritardi di registrazione degli eventi	Ritardi nelle registrazioni dei dati nella Fonte, comunicati dal fornitore	x
Ritardo - Ritardi di registrazione degli eventi	Differenza tra la data di registrazione degli eventi nella Fonte da parte del fornitore e la data di accadimento dell'evento nella popolazione	x
Ritardo - Ritardi di registrazione degli eventi	Nel caso di più forniture di un dataset riferite al tempo t ma estratte in momenti differenti: percentuale di oggetti (eventi o unità) registrati nell'ultimo dataset e mancanti nei precedenti	x
Stabilità delle variabili - Variazioni delle variabili o dei valori nel tempo	Metodi grafici (grafici a barre, scatter plot) per il confronto dei valori di specifiche variabili assunti dagli oggetti persistenti in differenti forniture della Fonte	x
Stabilità delle variabili - Variazioni delle variabili o dei valori nel tempo	Percentuale di oggetti con valore cambiato da t-1 a t (esclusi i missing) rispetto al totale di oggetti persistenti	x
Stabilità delle variabili - Variazioni delle variabili o dei valori nel tempo	Indici di associazione (V di Cramer) per variabili categoriali, o indici di correlazione per variabili numeriche, tra i valori di una stessa variabile al tempo t e al tempo t-1	x

Prospetto 3 segue **- Indicatori, informazioni/misure del report DATI**

INDICATORE	Informazione/Misura	Da implementare
Conformità	Numero di record per ciascun file che compone la fornitura, valori assoluti in serie storica	
Conformità	Percentuale dei campi valorizzati sul totale dei record valorizzati in serie storica per ogni variabile fornita	
Conformità	Frequenze percentuali delle modalità in serie storica per ogni variabile categorica	
Conformità	Modalità con decodifiche mancanti e corrispondente numero di record in serie storica per ogni variabile categorica	
Linkailità delle unità	Numero ed elenco delle variabili disponibili e utilizzate come chiavi di linkage per l'identificazione e la pseudonimizzazione delle unità	
Linkabilità delle unità	Percentuale di valori mancanti sul totale dei record per ogni variabile di linkage disponibile nella fornitura	
Accuratezza delle unità/monitoraggio del processo di integrazione	Serie storica della distribuzione percentuale di record per Tipo di abbinamento in SIM al momento del record linkage (ultimi due anni)	
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con Individui che non si sono abbinati in SIM al momento del record linkage e a cui è stato associato un nuovo Codice Individuo	
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con Individui che si sono abbinati in SIM al momento del record linkage	
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con Individui abbinati in SIM al momento del record linkage e persistenti nell'archivio	
Accuratezza delle unità/monitoraggio del processo di integrazione	Grafico in serie storica della percentuale di record con Individui abbinati in SIM al momento del record linkage e non persistenti nell'archivio (nuovi abbinati)	
Accuratezza delle unità/monitoraggio del processo di integrazione	Misure di variabilità dei tassi di link per sottopopolazioni	x
Accuratezza delle unità	Indicatori di qualità del record linkage	x
Accuratezza delle unità	Percentuale di record/oggetti che hanno un valore dell'identificativo errato (incompatibile con la definizione teorica della sintassi dell'identificativo o incoerente con quello presente in una lista di riferimento)	x
Accuratezza delle unità	Percentuale di record non autentici dichiarata dall'ente titolare della Fonte	x
Accuratezza delle unità	Percentuale di oggetti coinvolti in relazioni non logiche con altri oggetti o eventi	x
Accuratezza delle unità	Percentuale di oggetti coinvolti in relazioni ambigue ma non necessariamente errate con altri oggetti	x

Prospetto 3 segue **- Indicatori, informazioni/misure del report DATI**

INDICATORE	Informazione/Misura	Da implementare
Accuratezza delle unità	Ridondanza - Presenza di registrazioni multiple degli oggetti: percentuale di oggetti duplicati nell'archivio (con lo stesso codice identificativo)	x
Accuratezza delle unità	Ridondanza - Presenza di registrazioni multiple degli oggetti: percentuale di oggetti duplicati nell'archivio con gli stessi valori per un insieme di variabili	x
Accuratezza delle unità	Ridondanza - Presenza di registrazioni multiple degli oggetti: percentuale di oggetti duplicati nell'archivio con gli stessi valori per tutte le variabili	x
Accuratezza delle variabili	Errori di misura - Presenza di errori di misura sui valori delle variabili segnalati dal titolare della Fonte	x
Accuratezza delle variabili	Percentuale di oggetti con valori inconsistenti su una variabile (fuori range)	x
Accuratezza delle variabili	Percentuale di oggetti con combinazioni di valori non logiche su due o più variabili	x
Accuratezza delle variabili	Percentuale di oggetti con valori ambigui ma non necessariamente errati su una variabile	x
Accuratezza delle variabili	Percentuale di oggetti con combinazioni di valori ambigue ma non necessariamente errate su due o più variabili	x
Accuratezza delle variabili	Valori imputati - Presenza di valori derivanti da procedure di imputazione effettuate nel dataset dal fornitore dei dati per ciascuna variabile	x
Completezza (unità)	Sottocopertura – Oggetti target mancanti nel data set	x
Completezza (unità)	Sovracopertura - Presenza di oggetti non-target nel dataset	x
Completezza (unità)	Selectivity - Copertura per sottopopolazioni statistiche: Metodi statistici (distanze) per confrontare le distribuzioni degli oggetti nell'archivio e nella popolazione di riferimento rispetto ad una o più variabili di stratificazione	x
Completezza (variabili)	Valori mancanti - Assenza di valori per variabili di interesse : percentuale di oggetti con valore mancante per una particolare variabile	
Completezza (variabili)	Valori mancanti - Assenza di valori per variabili di interesse : percentuale di oggetti con tutti valori mancanti per un sottoinsieme di variabili	x

GLOSSARIO QRCA

Il Glossario bilingue, italiano e inglese, della QRCA comprende l'elenco dei termini utilizzati in Istat per la gestione e l'uso dei dati amministrativi. I lemmi che non hanno un corrispettivo in inglese sono stati specificatamente adottati in Istat.

Archivio amministrativo
Struttura del dataset amministrativo definita come sottoinsieme di una Fonte amministrativa.

Basi legali per l'utilizzo di dati amministrativi a scopi statistici
L'insieme delle disposizioni giuridiche che consentano all'Istituto Nazionale di Statistica di accedere ai dati amministrativi e di utilizzarli per la produzione di statistiche ufficiali e stabilire i termini e le condizioni alle quali tale uso è soggetto.
Legal Basis of the use of administrative data for statistical purposes: The whole set of legal provisions enabling an NSI to access administrative data and to use them for producing official statistics, and setting out the limits and conditions to which such use is subject.
Fonte: ESSnet Admin Data, WP1 (2013).

Completezza di un Dataset amministrativo
In termini di variabili, la misura in cui un *dataset amministrativo* contiene i valori dovuti delle variabili per le unità di una data *popolazione target*. In termini di *unità statistiche*, incompletezza in termini di unità è sinonimo di sotto-*copertura*.
Completeness of an Administrative Dataset: In terms of variables, the extent to which an *administrative dataset* contains the relevant items for the units of a given target population. Incompleteness in terms of units is synonymous with under-coverage.
Fonte: Adattato dalle definizioni di: BLUE-ETS Project, Deliverable 4.2 (2010); CODED, term extension "Statistical Concept"; SDMX Vocabulary; UNdata Glossary.

Copertura di un Dataset amministrativo
La capacità di un *dataset amministrativo* di contenere le unità *della popolazione target*. Misure di copertura normalmente si riferiscono al divario tra la *popolazione target* e l'insieme delle unità incluse nel dataset amministrativo disponibile. L'errore determinato dalla presenza, nel set di dati amministrativi a disposizione, di unità non appartenenti alla *popolazione target* è definito come "sovra-copertura", mentre l'assenza di unità della *popolazione target* nel dataset amministrativo disponibile, viene indicato come "sotto copertura".
Coverage of an Administrative Dataset: The extent to which an administrative dataset overlaps a given target population. Measures of coverage normally refer to the gap between the target population and the set of units included in the available dataset. The error due to units included in the available dataset but not belonging to the target population is often referred to as "over-coverage", while the error due to units belonging to the target population but not included in the available dataset is referred to as "under-coverage".
Fonte: ESSnet Admin Data, WP1 (2013).

Dataset amministrativo
Un insieme strutturato di dati estratti da una *Fonte amministrativa*, prima di qualsiasi trattamento o validazione da parte degli Istituti Nazionale di Statistica.
Administrative Dataset: Any organised set of data extracted from an *administrative source*, before any processing or validation by the NSIs.
Fonte: ESSnet Admin Data, WP1 (2013).

Dati amministrativi
Dati derivati da una *Fonte amministrativa*, prima di ogni processo di validazione da parte dell'Istituto nazionale di statistica.
Administrative Data: The data derived from an *administrative source*, before any processing or validation by the NSIs.
Fonte: Adattato dalle definizioni di: CODED, term extension "Statistical Concept"; SDMX Vocabulary; UNdata Glossary.

Dati amministrativi i un senso più ampio
Dati non raccolti principalmente per scopi statistici.

Dati prodotti a scopo commerciale
Dati prodotti da soggetti privati con lo scopo di essere venduti sul mercato.

Dati prodotti per finalità gestionali
Dati derivanti da attività gestionali condotte da soggetti pubblici o privati e che contengono informazioni utilizzabili a fini statistici. Si differenziano dai dati provenienti da Fonte amministrativa poiché non derivano dall'attuazione di uno o più regolamenti amministrativi.

Fonte amministrativa
Un insieme di dati raccolti e mantenuti per l'attuazione di uno o più *regolamenti amministrativi*. In un senso più ampio, qualsiasi fonte di dati che contiene informazioni che non sono raccolte principalmente per scopi statistici.
Administrative Source: A data holding containing information collected and maintained for the purpose of implementing one or more *administrative regulations*. In a wider sense, any data source containing information that is not primarily collected for statistical purposes.
Fonte: Adattato dalle definizioni di: CODED, term extension "Statistical Concept"; SDMX Vocabulary; UNdata Glossary[1].

Fornitore di dati amministrativi
Chi è incaricato dal *titolare dei dati amministrativi* a fornire i propri dati all'Istituto Nazionale di Statistica, per legge o in virtù di uno specifico accordo[2].
Administrative Data Provider: The *administrative data holder* who is bound to provide their data to the NSI, by law or by virtue of a specific agreement.
Fonte: ESSnet Admin Data, WP1 (2013).

1 Fonte originaria SDMX (2009).
2 La traduzione modifica la versione originale per dare maggiore chiarezza al termine.

Fornitura di dati amministrativi
Uno o più dataset amministrativi ricevuti dall'Istituto nazionale di statistica da parte del *Fornitore di dati amministrativi* in seguito ad una specifica richiesta. I data set della fornitura sono caratterizzati da un preciso riferimento temporale, inoltre ad essi può anche essere associata una data di estrazione dalla Fonte amministrativa o più generalmente una data di creazione. Nel caso in cui una stessa fornitura venga prodotta in tempi diversi, si può parlare di "Fornitura di dati provvisori" o "Fornitura di dati definitivi" o di "Aggiornamento dei dati della fornitura", ovviamente sempre con la specificazione della data di creazione. Un attributo della fornitura è la data di consegna.

Integrazione di dati
Il processo di combinazione di dati provenienti da due o più fonti per la produzione di dati statistici. L'integrazione dei dati può avvenire a livello micro, in questo caso viene denominata *matching* o al livello macro.
Data Integration: The process of combining data from two or more sources to produce statistical outputs. Data integration can be at the micro-level, where it is often referred to as *matching*, or at the macro-level.
Fonte: Definizione trovata in: CODED, term extension "Statistical Concept"; SDMX Vocabulary; UNdata Glossary.

Popolazione amministrativa
L'insieme di unità che una *Fonte amministrativa* è preposta a coprire, come definito dal relativo *regolamento amministrativo*. Questa popolazione può o non può corrispondere esattamente a una data *popolazione target*.
Administrative Population: The set of units that an *administrative source* is meant to cover, as defined by the relevant *administrative regulation*. This population may or may not correspond exactly to a given *target population*.
Fonte: ESSnet Admin Data, WP1 (2013).

Popolazione target
L'insieme delle unità per le quali si definiscono le statistiche e le stime richieste.
Target Population: The set of units about which information is wanted and estimates are required.
Fonte: Adattato dalle definizioni di: CODED, term extension "Statistical Concept"; OECD Glossary; SDMX Vocabulary; UNdata Glossary.

Regolamento amministrativo
Una serie di indicazioni dettagliate aventi vigore di legge, sviluppate per applicare una norma (come decreti, circolari e altre disposizioni simili). Solitamente la norma è rivolta ad una popolazione definita di persone fisiche e/o giuridiche, che è tenuta ad osservarla.
Administrative Regulation: A set of detailed directions having force of law, developed to put a policy into practice (such as decrees, ordinances, and other similar provisions). It is normally addressed to a designated population of natural and/or juridical persons, which are bound to observe it.
Fonte: ESSnet Admin Data, WP1 (2013).

Titolare della Fonte amministrativa
L'unità organizzativa titolare della *Fonte amministrativa*.
Administrative Source Holder: The organisational unit holding an *administrative source*.
Fonte: ESSnet Admin Data, WP1 (2013).

Unità amministrative
Con riferimento all'utilizzo dei *dati amministrativi* a fini statistici, le unità per le quali i *dati amministrativi* sono registrati. Queste unità possono o non possono essere uguali a quelle richieste per la produzione dell'output statistico (definite *unità statistiche*).
Administrative Units: With reference to the use of *administrative data* for statistical purposes, the units for which *administrative data* are recorded. These units may or may not be the same as those required for the statistical output (which are referred to as *statistical units*).
Fonte: ESSnet Admin Data, WP1 (2013).

Variabili identificative
Nell'ambito dei dati amministrativi, questo termine è usato per le variabili che permettono l'identificazione diretta delle *unità amministrative* e che sono necessarie per la raccolta, il controllo e l'integrazione dei dati, ma non sono successivamente utilizzate per l'elaborazione di risultati statistici (per esempio: Cognome, Nome, Partita IVA, Codice fiscale, Indirizzo, ecc.)
Identification Variables: Within the Administrative Data, this term is used to refer to those variables that allow direct identification of the *administrative unit* and which are needed for the collection, checking and *matching* of the data, but are not subsequently used for drawing up statistical results (for example: Surname, VAT number, Tax number, Address, etc.).
Fonte: Adattato dalle definizioni di: CODED, term extension "Statistical Methodologies"; OECD Glossary.

RIFERIMENTI BIBLIOGRAFICI

Ambroselli, S. 2015. "I codici identificativi univoci all'interno del SIM (Sistema Integrato di Microdati)". *Istat working papers*, N. 5/2015. Roma: Istat. https://www.istat.it/it/archivio/156101.

Calabria, M., L. Cappai, G. Di Bella, G. Petraccone, M. Porcelli, R. Rosati, G. Rotondi, e S. Spirito. 2018. "Interoperabilità dei sistemi IT per la gestione e la documentazione dei dati amministrativi". Poster presentato alla XIII Conferenza Nazionale di Statistica, *Dall'incertezza alla decisione consapevole: un percorso da fare insieme*. Roma, 4-6 luglio 2018.

Cerroni, F., G. Di Bella, and L. Galiè. 2014. "Evaluating administrative data quality as input of the statistical production process". *Rivista di statistica ufficiale*, N. 1-2/2014: 117-146. Roma: Istat. https://www.istat.it/it/archivio/136835.

Crescenzi, F., and F. Lipizzi. 2020. "The integration of geographic and territorial data sources into the base register of territorial and geographical entities". *Statistical Journal of the IAOS*, Volume 36, N. 1: 143–149.

Daas, P.J.H., S.J.L. Ossen, R.J.W.M. Vis-Visschers, and J. Arends-Tóth. 2009. "Checklist for the Quality evaluation of AD Sources". *Discussion paper (09042)*. The Hague, The Netherlands: Statistics Netherlands.

Daas, P.J.H., and S.J.L. Ossen. 2011a. "Metadata Quality Evaluation of Secondary Data Sources". *International Journal for Quality Research*, Volume 5, N. 2: 57-66.

Daas, P.J.H., S.J.L. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren. 2011b. "List of quality groups and indicators identified for administrative data sources". *Deliverable 4.1 of Workpackage 4 of the BLUE-ETS Project*.

Daas, P.J.H., S. J. L. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren, and B. Wallgren. 2011c. "Reports on methods preferred for the quality indicators of administrative data sources". *Deliverable 4.2 of Workpackage 4 of the BLUE-ETS Project*.

Di Bella, G. 2018. "L'interoperabilità conviene: documentare la qualità dei dati amministrativi utilizzati a scopi statistici in Istat". In *Atti della XIII Conferenza Nazionale di Statistica. Dall'incertezza alla decisione consapevole: un percorso da fare insieme*. Roma, 4-6 luglio 2018. Roma: Istat. https://www.istat.it/it/archivio/250234.

Drovandi, G., P. Giacomi, M. Giacummo, and E. Sibilio. 2017. "ARCAM: reengineering of Admin Data acquisition". Presentation at *New Techniques and Technologies for Statistics - NTTS 2017*. Brussels, 13-17 March 2017.

ESSnet Admin Data. 2013. "Admin Data Glossary. Definitions adopted for certain terms related to the use of administrative data for producing business statistics". *WP1 - Deliverable 1.1*.

Eurostat. 2017. *Codice delle statistiche europee. Per le autorità statistiche nazionali ed Eurostat (autorità statistica dell'UE). Adottato dal Comitato del sistema statistico europeo, 16 novembre 2017*. Lussemburgo: Ufficio delle pubblicazioni dell'Unione Europea.

Garante per la Protezione dei Dati Personali - GPDP. 2020. *Parere sullo schema di Programma statistico nazionale 2017-2019 - Aggiornamento 2019 - 13 febbraio 2020*. Roma: GPDP.

Garofalo, G. 2016. "Il Sistema dei registri come strumento di integrazione e miglioramento della qualità dei processi statistici". In *Atti della XII Conferenza Nazionale di Statistica. Più forza ai dati: un valore per il Paese*. Roma, 22-24 giugno 2016. Roma: Istat. https://www.istat.it/it/archivio/212411.

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ". *Journal of Official Statistics - JOS*, Volume 33, Issue 2: 477-511.

Runci, M.C., G. Di Bella, e L. Galiè. 2016. "Il sistema di integrazione dei dati amministrativi in Istat". *Istat working papers*, N. 18/2016. https://www.istat.it/it/archivio/193056.

Signore, M., M. Scanu, and G. Brancato. 2015. "Statistical Metadata: A Unified Approach to Management and Dissemination". *Journal of Official Statistics - JOS*, Volume 31, Issue 2: 325-347.

United Nations Economic Commission for Europe - UNECE, on behalf of the international statistical community. 2019. *Generic Statistical Business Process Model - GSBPM (Version 5.1, January 2019)*. Geneva, Switzerland: UNECE.

United Nations Economic Commission for Europe - UNECE. 2009. *Common Metadata Framework, Part A - Statistical Metadata in a Corporate Context: A Guide for Managers*. Geneva, Switzerland: UNECE.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration". *Statistica Neerlandica*, Volume 66, Issue 1: 41-63.