istat
working
papers

N.7
2023

Il monitoraggio e la valutazione della qualità del Sistema Integrato dei Registri

Monitoring and Evaluating the Quality of the Integrated System of Registers

Gabriele Ascari, Giovanna Brancato, Andrea Bruni, Stefano Daddi, Grazia Di Bella, Sara Giavante, Fabiana Rocci, Roberto Sanzo, Anna Maria Sgamba, Giorgia Simeoni, Simona Spirito

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.7
2023

Il monitoraggio e la valutazione della qualità del Sistema Integrato dei Registri

Monitoring and Evaluating the Quality of the Integrated System of Registers

Gabriele Ascari, Giovanna Brancato, Andrea Bruni, Stefano Daddi, Grazia Di Bella, Sara Giavante, Fabiana Rocci, Roberto Sanzo, Anna Maria Sgamba, Giorgia Simeoni, Simona Spirito

Istat

Istat Working Papers

Il monitoraggio e la valutazione della qualità del Sistema Integrato dei Registri

(Monitoring and Evaluating the Quality of the Integrated System of Registers)

N. 7/2023

ISBN 978-88-458-2121-9

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Il monitoraggio e la valutazione della qualità del Sistema Integrato dei Registri

(Monitoring and Evaluating the Quality of the Integrated System of Registers)

Gabriele Ascari[1], Giovanna Brancato[1], Andrea Bruni[1], Stefano Daddi[1], Grazia Di Bella[1], Sara Giavante[1], Fabiana Rocci[1], Roberto Sanzo[1], Anna Maria Sgamba[1], Giorgia Simeoni[1], Simona Spirito[1]

Sommario

Negli ultimi anni l'Istat ha sviluppato il Sistema Integrato dei Registri (SIR), che risponde alla nuova tendenza orientata verso una produzione statistica sempre più basata sull'organizzazione sistematica delle informazioni da fonte amministrativa integrate, ove necessario, con dati da indagini dirette. Il SIR è organizzato in diverse tipologie di Registri Statistici: quattro registri base (individui, unità economiche, luoghi e attività) e una serie di registri satellite, che hanno l'obiettivo di archiviare altre informazioni secondo criteri specifici, che possono essere direttamente riconducibili alle unità presenti nei registri base (registri estesi) o ne rappresentano un ulteriore approfondimento di dominio (registri tematici). In tale contesto è nata la necessità di definire un sistema di misurazione e valutazione della qualità del SIR, sia relativamente ai suoi singoli oggetti, sia alle relazioni che ne intercorrono. Questo documento presenta le basi poste per tale sistema, come frutto delle attività di un gruppo di lavoro, della durata di 6 mesi, opportunamente istituito in Istat. L'approccio proposto si basa su quattro principali direttrici: l'armonizzazione della terminologia tecnica, l'aderenza agli standard internazionali esistenti in materia, l'interoperabilità con sistemi sui metadati e la qualità già esistenti in Istat, l'ancoraggio ai processi di produzione sottostanti i registri del SIR. I principali risultati di questa attività sono: la sistematizzazione delle definizioni; l'identificazione di indicatori di qualità integrati in un modello standard di documentazione dei metadati; l'esplorazione di metodologie per la definizione di indicatori di qualità lungo tutto il processo produttivo; la mappatura di alcuni processi di produzione delle statistiche derivate da registri del SIR. Infine, questo documento delinea anche alcune riflessioni per ulteriori integrazioni metodologiche e alcune proposte per gli sviluppi necessari all'implementazione definitiva di tale sistema, quali l'ampliamento con ulteriori indicatori di qualità, la sperimentazione degli indicatori su alcuni registri e l'identificazione di misure di qualità sintetiche.

Parole chiave: registri statistici; indicatori di qualità; metadati.

Abstract

Over the last few years, Istat has developed the Integrated System of Registers (SIR), which addresses the new trend towards statistics produced from administrative data structured in statistical registers, and integrated with survey data, when necessary. The SIR is organised into base registers

1 Gabriele Ascari (gabascari@istat.it); Giovanna Brancato (brancato@istat.it); Andrea Bruni (anbruni@istat.it); Stefano Daddi (daddi@istat.it); Grazia Di Bella (dibella@istat.it); Sara Giavante (giavante@istat.it); Fabiana Rocci (rocci@istat.it); Roberto Sanzo (sanzo@istat.it); Anna Maria Sgamba (sgamba@istat.it); Giorgia Simeoni (simeoni@istat.it); Simona Spirito (spirito@istat.it), Istituto Nazionale di Statistica – Istat.

I punti di vista espressi sono quelli degli autori e non riflettono necessariamente le opinioni ufficiali dell'Istituto Nazionale di Statistica - Istat.

Gli autori ringraziano i revisori anonimi (almeno due per ogni lavoro, coinvolti su base volontaria e gratuita, con un approccio di tipo double - anonymised) per i loro commenti e suggerimenti, che hanno migliorato la qualità di questo Istat working papers N. 7/2023.

(base populations, e.g. individuals) and satellite registers. The latter aims to organise the information according to specific criteria, which in some cases can be directly linked to the units of the base registers (extended registers) or, in other cases, can represent a focus on given domains (thematic registers). In this context, the need to define a system to measure and assess the quality of the SIR has arisen, concerning both its objects and the relationships among them. To this aim, a six-month working group was purposely set up at Istat. This document reports the foundations laid down for the quality system, achieved by the working group. The proposed approach is developed along four main lines: the harmonisation of the technical terminology, the adherence to existing international standards, the interoperability with Istat existing systems on metadata and quality, and the anchoring to SIR production processes. The main results are the systematisation of definitions; the identification of quality indicators integrated into a standard metadata documentation model; the investigation of methodologies for the definition of quality indicators throughout the production process and the mapping of some statistical production processes underlying the SIR registers. Finally, proposals and reflections for future developments aimed at the final implementation of the quality system are also reported, such as the development of further quality indicators, the testing of the indicators on some registers, and the identification of synthetic quality measures.

Keywords: Statistical Registers; Quality Indicators; Metadata.

Lista degli acronimi usati:

ARCAM: Sistema di acquisizione di archivi amministrativi e dati di indagine;

ASIA: Registro statistico delle imprese attive;

Frame-SBS: Registro esteso delle principali variabili economiche delle imprese;

GdL: Gruppo di Lavoro;

GSBPM: Generic Statistical Business Process Model;

GSIM: Generic Statistical Information Model;

INS: Istituto/i Nazionale/i di Statistica;

Psn: Programma Statistico Nazionale;

QRCA: Quality Report Card dei dati Amministrativi;

RACLI: Registro annuale su retribuzioni, ore e costo del lavoro individuale;

RBI: Registro base degli individui, delle famiglie e delle convivenze;

RS: Registro Statistico;

RSB: Registro Statistico di Base;

RSBL: Registro Statistico di base dei luoghi;

RSE: Registro Statistico Esteso;

RST: Registro Statistico Tematico;

SIDI-SIQual: Sistema Informativo sulla Qualità dei processi statistici (SIDI: sistema di gestione, SIQual: Sistema di navigazione);

SIM: Sistema Integrato di Microdati;

SIMS: Single Integrated Metadata Structure;

SIR: Sistema Integrato dei Registri;

SIRIL: Sistema Integrato dei dati delle RILevazioni;

SUM: Sistema Unitario dei Metadati.

Indice

Pag.

Introduzione[2]

Il presente lavoro pone le basi per la costruzione di un impianto per il monitoraggio della qualità del Sistema Integrato dei Registri Statistici (SIR), sia in relazione ai processi alla base dei registri statistici che lo compongono sia delle statistiche prodotte a partire da esso. I risultati descritti nel documento sono il frutto delle attività svolte nell'ambito di un Gruppo di Lavoro (GdL) costituito all'Istat[3] per dare risposta all'obiettivo prefissato.

Per quanto concerne gli aspetti di contesto, va ricordato che il progetto di realizzazione del SIR è stato formalmente introdotto nell'ambito dei lavori del programma di modernizzazione dell'Istat, come uno dei tre capisaldi del processo di rinnovamento avviato nel 2016 (Istat, 2016*c*).

L'obiettivo del progetto SIR è il disegno e la costruzione di un sistema integrato di registri statistici, che rappresenti una struttura informativa unica e coerente a supporto dei diversi *output* della statistica ufficiale. Il risultato è ottenibile ponendo a sistema, concentrando e integrando, dati derivanti da una pluralità di fonti, in particolar modo incrementando l'uso di dati amministrativi, originati per finalità non statistiche da enti diversi dall'Istat, eventualmente combinati con dati da indagini dirette Istat.

Il progetto SIR costituisce un'innovazione strutturale considerevole nel modello di produzione statistica dell'Istat, che negli ultimi anni ha fatto registrare importanti avanzamenti nella costruzione dei singoli registri, del sistema disegnato per la loro integrazione e nell'utilizzo dei dati così prodotti.

L'organo dell'Istat formalmente incaricato del coordinamento strategico delle attività di realizzazione progressiva del SIR è il Comitato per la Gestione del sistema dei Registri (CGR)[4], chiamato a un impegno di particolare rilievo in termini di pianificazione strategica delle attività di costruzione e gestione del sistema quindi di definizione, realizzazione e monitoraggio nel perseguimento degli obiettivi del progetto stesso.

Nel corso dello sviluppo di tale sistema, ha preso corpo l'esigenza di definire un impianto che permetta il monitoraggio e la valutazione della qualità dei nuovi processi in essere e che rappresenti uno strumento utile ad affrontare i temi e le problematiche a essi connessi.

A tal fine, è stato costituito un GdL formato da professionalità trasversali al tema, con il compito di rilasciare il disegno di un *framework* che rappresentasse l'infrastruttura indispensabile per i metadati referenziali e strutturali, utile alla documentazione dei nuovi processi alla base dell'aggiornamento e gestione a regime dei registri statistici. Requisito del *framework* è di essere coerente con i princìpi già adottati in Istat e consolidati per altre tipologie di processo e armonizzato con gli standard concordati a livello internazionale.

In questo documento, quindi, sono descritte le varie attività portate avanti nell'ambito del GdL, che inizialmente ha teso a contestualizzare e focalizzare esattamente le esigenze

2 Il lavoro è frutto della collaborazione tra gli autori. Tuttavia l'introduzione e il paragrafo 1 sono da attribuirsi a Brancato e Rocci; il paragrafo 2 ad Ascari, Giavante, Rocci e Sgamba; i paragrafi 3 e 5 a Brancato, Bruni, Daddi, Rocci, Sanzo e Simeoni. Per il paragrafo 4, l'introduzione a Brancato, il sottoparagrafo 4.1. a Di Bella e il sottoparagrafo 4.2 a Brancato, Di Bella e Spirito. Per il paragrafo 6, i sottoparagrafi 6.1 e 6.2.1 sono da attribuirsi ad Ascari, i sottoparagrafi 6.2.2 e 6.3 a Daddi, e i sottoparagrafi 6.2 e 6.2.3 a Giavante; il paragrafo 7 a Brancato e Rocci.

3 Delibere DOP/755/2019 del 10/07/2019 e DOP/1073/2019 del 11/10/2019.

4 Delibere D08 20 DGEN 07/02/2017 e DOP 258 PRES 01/03/2019.

conoscitive e implementative del *framework* di qualità e individuare le innovazioni necessarie per rispondere alle peculiarità innescate dai nuovi processi.

Il GdL ha sviluppato le attività secondo diverse direttrici che hanno considerato: l'analisi degli standard condivisi tra i vari Istituti Nazionali di Statistica (INS) e le organizzazioni internazionali maggiormente coinvolte (Eurostat, UNECE); la descrizione della terminologia adottata nei contesti di produzione statistica *register-based* (definizioni, classificazioni, etc.); il riesame degli standard dei sistemi di documentazione e monitoraggio della qualità dell'Istat; la ricognizione dei registri adottati in alcune attività di produzione, tenendo anche conto del loro diverso grado di maturità di implementazione; infine, la definizione di una proposta strutturata di un *framework* di documentazione dei nuovi processi a supporto del monitoraggio e della valutazione della qualità. Questa proposta è in particolare declinata in uno degli *output* fondamentali del lavoro: l'insieme delle schede di modellazione per la documentazione e la definizione di indicatori di qualità dei processi del SIR.

L'ottica in cui il GdL ha lavorato è quella che assume l'implementazione a regime del SIR, tralasciando quindi tutte le attività di controllo proprie della progettazione del sistema, ma tenendo conto delle varie tipologie di registro che ne fanno parte, delle regole di interconnessione tra i registri e delle attività aggiornamento degli stessi.

Il documento è così organizzato. Il paragrafo 1 presenta il contesto generale da cui ha avuto origine l'esigenza di questo lavoro, gli obiettivi che hanno guidato il GdL e il perimetro delle attività.

Il paragrafo 2 tratta gli aspetti definitori e concettuali con un forte collegamento a quelli che sono i concetti adottati in ambito internazionale e come questi siano applicati in Istat.

Il paragrafo 3 presenta una sintesi degli aspetti più salienti della modellazione a livello internazionale sui metadati e gli indicatori di qualità, utilizzati per le finalità di questo lavoro. Si tratta di esperienze maturate nell'ambito del Sistema Statistico Europeo e in collaborazione con le istituzioni statistiche delle Nazioni Unite, in particolare l'UNECE (*SIMS*, *GSIM*, *GSBPM*).

Il paragrafo 4 entra nel dettaglio dell'acquisizione dei all'Istat, con particolare *focus* sugli aspetti operativi, di controllo e di monitoraggio dei dati amministrativi che rappresentano l'*input* centrale del SIR. Sono inoltre evidenziate le relazioni con i sistemi di documentazione e qualità già esistenti all'Istat.

Il paragrafo 5 presenta la proposta di modellazione degli indicatori di qualità (per i processi, per i prodotti e per le relazioni tra registri) integrata con i metadati e sviluppata considerando gli indicatori proposti in letteratura, quelli adottati all'Istat e i più recenti standard internazionali. Viene anche presentato un esempio di schematizzazione del processo del registro relativo ai conti economici delle imprese denominato Frame-SBS.

Il paragrafo 6 presenta uno studio sperimentale di costruzione di un sistema di indicatori definiti attraverso il ciclo di vita dei dati. Il lavoro rappresenta una possibile generalizzazione al Registro base degli individui, delle famiglie e delle convivenze (RBI) di un approccio adottato dall'Istituto Nazionale di Statistica Austriaco. La finalità è quella di individuare possibili spunti per ulteriori innovazioni sui criteri di misurazione e valutazione della qualità lungo tutto il ciclo di vita del dato.

Il documento si chiude con una disamina sintetica di quanto realizzato e delle attività da sviluppare a completamento del progetto (paragrafo 7).

1. Contesto e finalità del gruppo

1.1 Obiettivi del gruppo

Nell'ultimo decennio, i processi statistici dell'Istat sono stati oggetto di grandi innovazioni, partendo dall'acquisizione sistematizzata su larga scala di dati da fonte esterna, in primis dati amministrativi, fino alla formulazione dei nuovi processi multi-fonte. Inoltre, la definizione e la costruzione di un sistema integrato di registri statistici, con l'obiettivo di cambiare il paradigma della produzione della statistica ufficiale, rappresenta un importante *milestone* del processo di riorganizzazione nell'utilizzo efficiente del potenziale conoscitivo offerto dalle varie fonti esistenti su fenomeni di interesse.

A tal fine, sono state definite nuove metodologie per lavorare in situazioni di una molteplicità di fonti dettagliate e di elevata copertura, utilizzabili a fini statistici, incentivandone così l'utilizzo per lo sviluppo di singoli registri statistici e di un sistema integrato.

Ai fini della generale rappresentazione del SIR, è stata fondamentale un'analisi degli aspetti concettuali, definitori e classificatori attinenti a questo contesto. Questa è stata considerata necessaria e propedeutica alla definizione delle fasi standard del processo di costruzione di un registro statistico, considerato che, rispetto al tradizionale modello di produzione dell'Istat basato sulle indagini dirette, l'impiego dei registri statistici nei processi di produzione impone un riesame nella rappresentazione dei processi stessi.

La complessità del SIR ha richiesto la formulazione di obiettivi specifici, che permettessero di individuare gli elementi utili alla ricognizione prima e alla definizione successiva dei concetti e delle modalità di costruzione di un sistema di monitoraggio dei processi e degli *output* dei registri statistici.

In questa ottica, il GdL è stato formalizzato definendo i seguenti obiettivi:

a. analisi di quanto già disponibile in Istat rispetto alla documentazione e misurazione della qualità e di quanto necessario in un'ottica di generalizzazione;
b. definizione dei metadati e delle misurazioni funzionali alla documentazione, al monitoraggio e alla valutazione della qualità dei processi alla base dei registri statistici;
c. valutazione del sistema di documentazione tenendo conto della natura del registro: base, tematico o esteso;
d. analisi dei requisiti architetturali e tecnici per l'implementazione del sistema di documentazione e misurazione.

Allo scopo di sviluppare un modello che riflettesse concretamente le buone pratiche adottate all'Istat, sono stati selezionati tre registri per analizzare in dettaglio i loro processi e i relativi controlli di qualità. I registri inclusi nelle analisi sono stati specificatamente:

- Registro statistico delle imprese attive (ASIA);
- Registro esteso delle principali variabili economiche delle imprese (Frame-SBS);
- Registro base degli individui, famiglie e convivenze (RBI).

Essi sono rappresentativi sia delle unità economiche sia di quelle sociali e riflettono livelli di maturità diversi essendo i primi due già ampiamente consolidati e dotati di un loro sistema di monitoraggio e il terzo costruito nell'ambito delle più recenti innovazioni introdotte con la strategia di modernizzazione nel 2016.

L'ambito del lavoro del gruppo è stato circoscritto all'aggiornamento dei registri del SIR attraverso le seguenti fasi:

- acquisizione e/o trasmissione interna dei dati (di fonte amministrativa, da indagine, da altro registro);
- valutazione della qualità delle fonti di dati trasmessi per le finalità del registro; trasformazioni dei dati delle fonti;
- codifica e classificazione dei dati delle fonti o dei dati integrati; integrazione tra le fonti; controllo e correzione dei dati delle singole fonti o dei dati integrati;
- derivazione di nuove variabili e unità. Sono quindi stati presi in considerazione alcuni sottoprocessi di due fasi del modello *GSBPM*.

Il primo passo ha riguardato la ricognizione della letteratura sui concetti e criteri di misurazione della qualità e dei sistemi già implementati, perché rappresentano un'utile base per un'analisi critica di ciò che è applicabile direttamente o con modifiche e personalizzazioni per tener conto delle caratteristiche innovative del contesto specifico. La documentazione dei processi, degli *output* e della loro qualità, oltre a essere utile agli utenti finali dei dati, è indispensabile per le finalità dei sistemi di monitoraggio interni all'Istituto.

All'interno delle varie fasi di processo e seguendo la loro rappresentazione mediante metadati si arriva a individuare le esigenze conoscitive che gli indicatori dovrebbero descrivere, quale informazione dovrebbero rilasciare e quali dovrebbero essere gli *input* necessari alla loro elaborazione.

Da un punto di vista metodologico, un contributo importante per la definizione degli indicatori deriva da alcuni lavori in letteratura, il cui studio ha fornito degli elementi interpretativi delle problematiche per arrivare a una loro adeguata rappresentazione e comunicazione agli utenti. In particolare, due ambiti di ricerca sono stati sottoposti ad analisi per i propri schemi interpretativi. Il primo riguarda un approccio orientato alla valutazione in un'ottica *Total Survey Error* (TSE), con l'individuazione delle sue fonti di errore applicata a processi basati su integrazione di dati (Zhang, 2012; Zabala *et al*., 2013; Reid *et al*., 2017; Luzi *et al*., 2018). Il punto di forza di questo approccio è l'individuazione della natura a due fasi dei processi statistici che utilizzano fonti amministrative, dove la prima fase è rappresentata dall'analisi delle singole fonti, mentre nella seconda fase ha luogo l'integrazione e la trasformazione dei dati in informazione statistica. Questo schema è già stato recepito in Istat nei suoi concetti generali, per esempio dalla stessa QRCA[5] (*Quality Report Card* dei dati Amministrativi) che documenta i metadati e la qualità della prima fase di acquisizione (Venturi e Di Bella, 2017). Nel nostro lavoro l'obiettivo è quello di standardizzare gli aspetti che si presentano nella seconda fase, dalla trasformazione in dati con specifiche finalità statistiche.

5 Portale di documentazione dei dati amministrativi acquisiti dall'Istat a fini statistici raggiungibile dalla intranet dell'Istat all'indirizzo https://qrca.istat.it/QRCA/ (cfr. sottoparagrafo 4.2).

Il secondo ambito di ricerca preso in considerazione è orientato a sviluppare un insieme di misure sintetiche della qualità lungo tutto il processo produttivo del dato statistico e un loro modo per sintetizzarle in una valutazione complessiva del registro (Asamer *et al.*, 2016).

La complessità progettuale e organizzativa dei lavori di realizzazione del SIR, con forti interazioni tra i diversi settori e strutture dell'Istat, ha previsto a livello operativo un'articolazione delle attività che ha specificato, seppur in un quadro di cooperazione flessibile, i ruoli delle macrostrutture dell'Istat per i basilari macroprocessi individuati. Questo ha comportato l'esigenza di articolare il gruppo stesso in professionalità e compiti appartenenti alle varie strutture dell'Istat, che in modo diverso concorrono al compimento dei singoli registri e del sistema.

Tale impostazione dei lavori, fortemente trasversale tra le varie direzioni centrali dell'Istat che a vario titolo concorrono alla implementazione dei processi alla base dei registri statistici, ha permesso di valutare tutte le esigenze conoscitive e gli aspetti sia metodologici sia operativi da documentare e monitorare. Infatti, l'obiettivo finale delle attività del GdL era di fornire ai responsabili dei registri statistici del SIR un insieme di misurazioni standard completo e coerente, che consentisse loro il monitoraggio della qualità sia inizialmente in fase di disegno a supporto del processo decisionale, sia durante le fasi di produzione dei registri, sia a fine processo per finalità documentative e valutative.

1.2 Sistemi di riferimento per la documentazione e il monitoraggio della qualità

Il presidio della qualità dei processi e dei prodotti statistici rientra tra i compiti istituzionali degli INS. Tale funzione è ancora più importante quando si introducono innovazioni che possono avere un impatto rilevante a vari livelli.

Anche l'Istat ha un presidio di lungo periodo sulla qualità, ampiamente consolidato nell'ambito dei processi che possono essere indicati come "tradizionali", ovvero basati sull'utilizzo di dati provenienti da indagine diretta, sia essa campionaria sia esaustiva. Per quanto riguarda le innovazioni, a partire dall'utilizzo di dati amministrativi, l'Istat ha prontamente sviluppato e consolidato un sistema per il monitoraggio della loro qualità come dati di *input* (QRCA) (Di Bella, 2021) e ha posto le basi per la valutazione dei processi che li utilizzano (Brancato *et al.*, 2016).

Per lo sviluppo del *framework* di qualità sulle attività successive all'acquisizione, il punto di partenza è stato il riferimento agli standard consolidati in campo internazionale e già alla base dei sistemi esistenti in Istat. Tali standard suggeriscono di procedere *in primis* con la rappresentazione delle fasi di un processo, per fornire la sua documentazione in termini di metadati referenziali e strutturali. In particolare, il *Generic Statistical Business Process Model* (*GSBPM*) versione 5.1 (UNECE, 2019) è lo schema che risponde a un'esigenza di classificazione e armonizzazione delle diverse fasi dei processi di produzione statistici messi in atto dagli INS (Istat, 2016*a*) e il *Generic Statistical Information Model* (*GSIM*) versione 1.2 (UNECE, 2021) descrive per ognuna di queste attività gli *input* e gli *output* secondo i criteri di classificazione dei metadati referenziali e strutturali.

In questa ottica, un ruolo importante è rappresentato da sistema SIDI-SIQual[6] (Sistema Informativo sulla Qualità dei processi statistici), già implementato in Istat, che documenta i metadati referenziali e gli indicatori standard di qualità per il monitoraggio dei processi e dei prodotti statistici, in coerenza con le definizioni adottate nell'ambito del Sistema Statistico Europeo (Eurostat, 2003). Il sistema, nato per le indagini dirette accoglie a oggi anche la documentazione dei registri statistici, anche se non secondo gli standard più recenti. Esso contiene:

- l'elenco dei processi statistici dell'Istat con il loro codice nel Programma statistico nazionale (Psn);
- i regolamenti sottostanti ai processi;
- le informazioni anagrafiche dei processi (descrizione, varie periodicità, anno di prima produzione);
- i metadati di processo, ossia tutte le fasi e i sottoprocessi, sia a regime sia in caso di ristrutturazione, organizzati in modo gerarchico e con un ampio dettaglio su come queste siano svolte;
- i metadati sulla qualità ossia tutte le azioni di controllo della qualità adottate per prevenire, controllare in corso d'opera e valutare a posteriore gli errori;
- gli indicatori standard di qualità.

La documentazione già disponibile in SIDI-SIQual è stata utilizzata come punto di partenza per l'identificazione delle nuove esigenze in termini di metadati referenziali e indicatori di qualità.

I lavori del GdL si sono sviluppati tracciando la vita del dato, in particolare di quello amministrativo, dalla sua origine, alla sua trasmissione da enti esterni, alla sua gestione centralizzata in Istat e al suo utilizzo nei vari processi, fino alla finale trasformazione in dato statistico. In particolare, lo studio del GdL si aggancia alla fase di gestione centralizzata dell'arrivo dei dati, pertanto le esigenze coperte da questo lavoro sono relative al disegno delle attività successive, sviluppate nei vari processi di produzione, per delinearne gli aspetti comuni e proporne una rappresentazione standardizzata. In sostanza, l'obiettivo è definirne una documentazione generalizzata, con l'adeguata flessibilità a seconda delle tipologie di registro e di processo, da cui dipendono i metodi e le problematiche.

Nell'ambito del SIR, è importante sottolineare come per la sua costruzione, ricoprano un ruolo fondamentale due sistemi presenti in Istat: il Sistema Integrato di Microdati (SIM) e il Sistema Unitario dei Metadati (SUM)[7]. Il SIM rappresenta l'integrazione logico-fisica di microdati amministrativi a partire da archivi di *input* provenienti da fonti amministrative e permette di identificare tutte le unità che entrano nel SIR mediante il codice univoco denominato codice SIM (Runci *et al.*, 2016). Il SUM Sistema Unitario dei Metadati - Metadati strutturali acquisisce e gestisce i metadati disponibili nei sistemi dell'Istat dalla fase di acquisizione alla fase di diffusione, secondo gli standard dell'Istat e gli standard

6 Il sistema di navigazione SIQual è accessibile dal sito esterno dell'Istat al seguente *link*: http://siqual.istat.it/SIQual/welcome.do. Nella versione *intranet* è al seguente *link*: https://siqual.intra.istat.it/SIQual/welcome.do. Il sistema di gestione SIDI è accessibile solo mediante credenziali.

7 Il sistema è accessibile dalla intranet dell'Istituto al seguente *link*: http://web3-int-svil.istat.it:8080/SUM/noSecure/login.action.

internazionali, garantendo nel SIR la coerenza delle classificazioni e dei metadati (Istat, 2017). Esso è dedicato alla descrizione del contenuto e del significato degli archivi di dati (ovvero sui metadati strutturali). Contiene: la definizione di strutture di dati micro e macro prodotte lungo qualsiasi programma statistico intrapreso dall'Istat; i concetti necessari per le definizioni delle strutture di micro e macro-dati.

SUM e SIDI-SIQual sono complementari per le informazioni sui metadati e sono concettualmente e tecnicamente integrati tra loro.

Infine, la fase di acquisizione dei dati, attraverso una gestione centralizzata, risulta strutturata e rilascia già un sistema di monitoraggio dei dati amministrativi in entrata in Istat, attraverso la già citata QRCA, che a sua volta rappresenta uno dei punti di aggancio dei lavori di tale gruppo.

2. Quadro definitorio e concettuale di riferimento per il SIR

Il SIR si basa su un insieme di concetti e definizioni che derivano dalla documentata esperienza internazionale sul graduale e sempre più consistente ricorso all'uso di dati amministrativi e alla costruzione degli *output* della statistica ufficiale utilizzando dati di registri statistici. Il processo di riorganizzazione dell'Istat avviato nel 2016 ha poi richiesto uno sforzo di sistematizzazione per recepire la letteratura internazionale integrandola e adattandola alla realtà interna e ai progetti in essere. È fondamentale quindi analizzare una serie di definizioni che permettano di chiarire i termini che sono alla base del SIR, nell'ottica di uniformare la terminologia e di esplicitare i concetti più utilizzati all'Istat per meglio comprendere il complesso sistema dei registri.

I punti di riferimento in letteratura sono i manuali rilasciati dai paesi del Nord Europa, che hanno una lunga tradizione nel campo (UNECE, 2007; Wallgren and Wallgren, 2014), e i vari documenti Istat che descrivono gli schemi adottati e riproposti adattandoli per definire il SIR (Istat, 2016*a*; Istat, 2016*b*, Istat, 2016*c*; Istat, 2017).

Il ricorso da parte degli INS all'impiego di dati amministrativi per fini di statistica ufficiale è relativamente recente. Il sistema di condivisone di tali dati coinvolge attori specifici (enti o istituzioni pubbliche) e nell'ultimo decennio si è sviluppato in modo molto veloce e strutturato. La relativa terminologia deve tener conto del fatto che i dati provenienti da fonti esterne nascono per finalità diverse da quelle prettamente statistiche e presuppongono definizioni delle misurazioni, modalità di raccolta e archiviazione dei dati proprie dell'ente interessato.

In generale un registro è definito come una raccolta sistematica di dati relativi a unità statistiche organizzata in modo tale che ne sia possibile l'aggiornamento. L'aggiornamento è inteso in senso ampio come il trattamento di informazioni identificabili allo scopo di stabilire, aggiornare, correggere o estendere il registro, vale a dire tenere traccia di eventuali cambiamenti nei dati descrivendo le unità e i loro attributi.

È importante sottolineare che gli INS hanno sviluppato metodi di archiviazione e di aggiornamento sistematici secondo dei criteri standard di classificazione degli attributi delle unità/oggetto dell'archivio, a seconda delle informazioni e della tipologia provvisoria o definitiva della fornitura. Inoltre, negli ultimi decenni è stato fatto un grande investimento da parte degli INS per coordinarsi con gli enti fornitori e sfruttare le informazioni legate ai dati amministrativi così archiviati, da qui l'importanza della conoscenza approfondita di tali basi dati, nei loro attributi e nella loro metadatazione.

Per questo motivo, ripercorriamo di seguito le definizioni e i criteri nel campo dell'archiviazione dei dati, per tenere presente le eventuali differenze tra quelli adottati da altri istituti e quelli invece peculiari degli INS (o degli enti del Sistan in generale). Infatti, per quanto l'informazione di base sia la stessa, l'obiettivo e le definizioni a cui devono rispondere cambiano gli attributi che i vari metodi di sistematizzazione dell'informazione adottano.

Il SIR stesso viene alimentato da un sistema centralizzato, previa armonizzazione e strutturazione delle fonti amministrative.

Nel sottoparagrafo 2.1 si chiarisce il concetto di dato amministrativo, alla base di ogni sistema fondato su registri, e si descrive il significato di Registro Amministrativo e Registro Statistico. Nel sottoparagrafo 2.2 vengono descritte le varie tipologie di Registro Statistico, in funzione delle componenti che lo caratterizzano. Infine nel sottoparagrafo 2.3 si esplicitano i tre livelli dell'integrazione che caratterizzano il SIR.

L'analisi approfondita delle definizioni adottate dalle strutture dell'Istat, che con diverso ruolo si occupano del SIR, ha evidenziato un certo livello di disallineamento interno e con la letteratura di riferimento, nel senso che la terminologia non è sempre utilizzata in modo univoco. Questa evidenza potrebbe suggerire la necessità di un ulteriore approfondimento e l'adozione di un unico glossario specifico e ufficiale dell'Istituto. Questo non era tra gli obiettivi definiti nella delibera di costituzione del gruppo e pertanto si è deciso di non perseguirlo durante le attività del gruppo stesso.

2.1 Concetti di base

Di seguito sono elencate le definizioni raccolte nei vari documenti dell'Istat. Alcune rispecchiano completamente i concetti presenti in letteratura, in altri casi sono state adattate al contesto interno dell'Istat e non trovano esatta corrispondenza con la letteratura. Quando ciò succede, sono riportate delle riflessioni utili a valutare l'opportunità di allineare tali situazioni.

Le definizioni individuate sono:

- Dato amministrativo: informazione raccolta e conservata da un'istituzione pubblica ai fini di controllo o di intervento nei confronti di singoli individui o entità di altro tipo (secondo regolamenti amministrativi). Per l'Istat si tratta di un dato proveniente da una fonte amministrativa non sottoposto a elaborazioni o validazioni da parte dell'Istat (Costanzo *et al.*, 2013).
- Fonte amministrativa: insieme di informazioni raccolte per scopi amministrativi e non statistici da parte dell'Ente titolare del dato amministrativo (Costanzo *et al.*, 2013).
- Archivio amministrativo: per l'Istat rappresenta il *dataset* amministrativo che comprende un sottoinsieme di informazioni di una fonte amministrativa (ad esempio l'Archivio MIUR - Archivio degli iscritti e delle iscrizioni universitarie è estratto dalla fonte amministrativa Anagrafe Nazionale degli Studenti del MIUR). Questa definizione trova ampio utilizzo presso tutte le strutture dell'Istituto (Glossario della QRCA in Di Bella, 2021). Nella letteratura sulla creazione di sistemi di registri statistici, nonché nel glossario recepito dall'Istat nello stesso ambito, tale definizione sembra coincidere con quella di Registro Amministrativo (UNECE, 2007; Wallgren and Wallgren, 2014), riportato di seguito.
- Registro Amministrativo (RA): raccolta sistematica strutturata, regolarmente aggiornata e autorizzata di dati e metadati (dati e metadati costituiscono le proprietà del registro) a livello di unità o evento (le unità e gli eventi rappresentano gli oggetti del registro) per una specifica popolazione, in conformità all'attuazione di determinate

finalità amministrative. Gli oggetti del registro sono esaustivi per la popolazione individuata dalle rispettive proprietà e sono identificati in maniera inequivocabile da un codice univoco. Gli oggetti del RA sono individuati da un preciso insieme di norme di tipo amministrativo e le rispettive proprietà derivano da classificazioni stabilite da regolamenti amministrativi, (Istat, 2016*a*; UNECE, 2007; Wallgren and Wallgren, 2014). Come descritto precedentemente questa definizione, sebbene diffusa a livello internazionale, non trova largo uso in Istituto.

- Registro Statistico (RS): raccolta sistematica strutturata, regolarmente aggiornata e autorizzata di dati e metadati (proprietà) a livello di unità o evento (oggetto) per una specifica popolazione effettuata esclusivamente per finalità di statistica ufficiale. Il RS è realizzato da un ente facente parte del Sistan. Gli oggetti del registro sono determinati da definizioni e classificazioni che derivano da criteri statistici, connessi alle esigenze della statistica ufficiale, e sono identificati in maniera inequivocabile da un codice univoco. I RS sono rappresentati da unità statistiche caratterizzate da variabili statistiche (Istat, 2016*a*). Il RS può essere creato e aggiornato a partire da una o più fonti anche combinate tra loro: registri amministrativi, registri statistici e indagini statistiche. Un RS può essere produttore di informazioni statistiche oppure costituire la base di riferimento a supporto di vari processi che producono informazioni statistiche (Wallgren and Wallgren, 2014).

2.2 Tipologie di Registri Statistici

In tutti gli schemi proposti per un sistema di registri è importante la loro distinzione per tipologia. Tale classificazione è rilevante perché le caratteristiche e gli attributi di un registro possono guidare o addirittura determinare la selezione delle informazioni necessarie per il loro monitoraggio.

Una ricognizione della letteratura ha evidenziato degli elementi comuni tra loro, ma non una completa uniformità. A partire delle definizioni rilasciate in Istat, sono stati analizzati i principali elementi che possono aiutare nella completa mappatura dei registri nell'ambito del SIR.

I tre elementi che principalmente guidano le definizioni delle differenti tipologie di registro sono: il tipo di informazione obiettivo, le caratteristiche delle unità statistiche su cui sono basati e il tipo del dato amministrativo utilizzato come *input* del processo.

Le diverse combinazioni di questi tre elementi determinano le diverse definizioni di tipologia di registro. La prima distinzione che ne scaturisce è tra Registri Base e Registri Satellite, a loro volta suddivisi tra Estesi e Tematici[8], per cui di seguito si formulano le seguenti definizioni.

Registri Statistici di Base (RSB): le informazioni obiettivo sono quelle necessarie perché le unità possano essere riconosciute come appartenenti a una specifica popolazione base della statistica ufficiale. Le informazioni che trattano sono di diverso tipo, e giocano un ruolo fondamentale quelle derivate direttamente da fonte amministrativa.

8 Nei vari schemi proposti in letteratura e all'Istat, sono sempre presenti i registri della tipologia "base". Invece la mappatura degli altri registri non è sempre adottata in modo uniforme.

In questo ambito l'Istat ha adottato la struttura proposta da Wallgren and Wallgren (2014), di definire come base 4 registri, di cui uno con peculiarità di carattere trasversale che racchiude le informazioni sulle relazioni tra gli altri (Istat, 2016*c*).

Nello specifico sono:

a. Registro base degli individui, delle Famiglie e delle Convivenze (RBI);
b. Registro delle unità economiche (imprese e istituzioni);
c. Registro statistico di base dei luoghi (RSBL);
d. Registro delle attività, relativo alle attività e agli eventi (ad esempio, lavoro o studio) che lega le unità appartenenti ai vari registri base.

I primi tre registri riguardano le unità delle popolazioni più rilevanti per la statistica ufficiale: individui, unità economiche e territorio e, le informazioni necessarie per individuare tali unità per le varie statistiche a cui devono servire. Il registro delle attività permette, invece, a partire dalle unità statistiche degli altri registri di costruire le relazioni che servono a legare tra loro le persone, le unità economiche e i luoghi.

Registri Statistici Satellite: hanno l'obiettivo di rilasciare altre variabili di tipo tematico, ad esempio, educazione, salute, sicurezza, reddito, etc., derivate, quando possibile, direttamente dalle fonti amministrative, oppure integrando in modo opportuno le informazioni dalle rilevazioni.

La loro successiva suddivisione dipende dalla tipologia di unità statistica alla loro base e dal tipo di dato di *input*. Quindi si hanno:

- Registri Statistici Estesi (RSE): le unità statistiche possono essere identificate tra quelle appartenenti a uno di quelli base. Le variabili che rilasciano riguardano informazioni solitamente relative a fenomeni specifici, spesso individuati da Regolamenti EU. Tra le tipologie di dati che integrano hanno un ruolo fondamentale quelle direttamente rilevate da una o più fonti amministrative. All'interno dell'Istat un esempio di registro di questa tipologia è il registro Frame-SBS, che sarà oggetto di ulteriori analisi all'interno di questo lavoro.
- Registri Statistici Tematici (RST): anche questi registri forniscono informazioni su fenomeni specifici, ma si differenziano dagli estesi per la presenza di almeno uno dei due concetti seguenti:
    - l'unità statistica fondamentale non è un'unità statistica di uno dei registri base, anche se a quelle deve poter essere riconducibile. Ciò li rende trasversali, perché costruiti a partire da informazioni formalizzate in base a unità statistiche differenti dagli individui o dalle unità economiche. Un esempio di registro tematico dell'Istat è il Registro Tematico del Lavoro (RTL), che è sviluppato a partire dall'unità statistica "posizione lavorativa";
    - i dati elaborati in *input* non provengono necessariamente da fonti amministrative pure, ma si possono basare anche su degli *output* di altri processi statistici all'interno del SIR. Quindi in quest'ultimo caso non si configurerebbe un uso diretto di fonti amministrative, bensì di dati già elaborati secondo criteri statistici.

2.3 Criteri del sistema di integrazione

Dal punto di vista operativo, il SIR è un ambiente informativo strutturato per essere adibito a vincolo e a supporto dei processi di produzione dell'Istat. Nel sistema, la gestione delle informazioni è governata da regole prestabilite finalizzate a garantire la coerenza e non ridondanza dei dati. Per esempio, per ogni variabile è stabilito quale registro abbia il compito della gestione e dell'aggiornamento, pur consentendone l'accesso agli altri registri (Istat, 2016*c*). Un sistema di integrazione è un insieme costituito da criteri, metodologie e processi che garantiscono la combinazione coerente degli oggetti e delle proprietà provenienti da più fonti informative. I criteri, le metodologie e i processi messi in atto assicurano tre diversi livelli di integrazione: concettuale, logico/fisica e statistica.

Nello specifico:

- integrazione concettuale: criteri, metodologie e processi che permettono di identificare i metadati (definizioni e classificazioni) di oggetti e proprietà contenuti nelle fonti informative e assicurano la coerenza tra oggetti e proprietà di una stessa fonte informativa nel tempo e tra oggetti e proprietà di più fonti informative;
- integrazione logico/fisica: criteri, metodologie e processi che permettono di identificare uno stesso oggetto presente in fonti informative diverse mediante un identificativo univoco e stabile nel tempo (ID) e che consentono di definire e identificare per ogni oggetto, sia le relazioni logiche e fisiche nel tempo e nello spazio tra le informazioni provenienti da fonti informative diverse, sia le relazioni logiche e fisiche tra oggetti che si riferiscono a una stessa tipologia di unità statistiche (ad esempio relazioni tra individui) e tra oggetti che afferiscono a differenti tipologie di unità statistiche (ad esempio relazioni tra individui e unità economiche);
- integrazione statistica: criteri, metodologie e processi che permettono di combinare oggetti e proprietà di più fonti informative in modo da poter identificare popolazioni e sottopopolazioni statistiche, integrare variabili, derivare variabili, rispettando vincoli di coerenza interna (tra le statistiche prodotte dal SIR) ed esterna (tra le statistiche del SIR e le statistiche prodotte dall'Istat). L'integrazione statistica permette, inoltre, di definire la coerenza tra oggetti e le loro proprietà desumibili da più fonti informative.

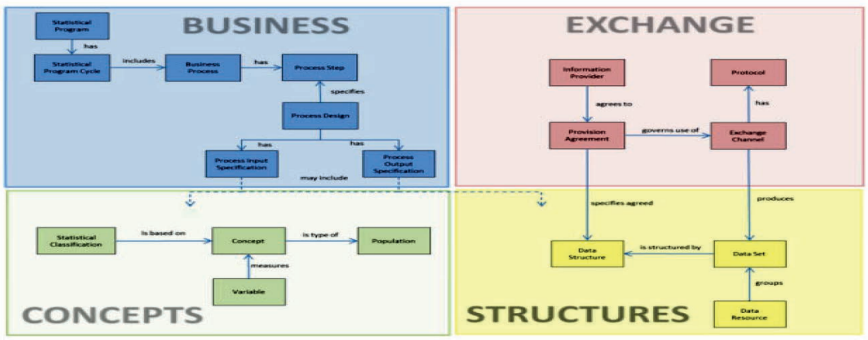
3. Modelli e standard di riferimento per i metadati e gli indicatori di qualità

Gli standard internazionali di riferimento relativi ai metadati presi in considerazione in questo lavoro sono *GSIM* e *GSBPM*. In particolare questo lavoro ha beneficiato delle attività parallele a livello internazionale che si stanno conducendo sull'integrazione tra i suddetti standard (UNECE, 2020).

Relativamente alle misure di qualità, è stato assunto come base di partenza il set di indicatori standard di qualità definiti a livello Eurostat e integrati nella struttura di metadati denominata *Single Integrated Metatada Structure* (Eurostat, 2020). Tuttavia, non si è rinunciato ad ampliare la gamma di misure da prendere in considerazione e sono stati valutati anche indicatori non compresi nello standard europeo.

GSIM modella oggetti informativi, come per esempio dati, metadati, regole, parametri e, più in generale, tutti quegli oggetti necessari alla produzione statistica. Questi sono raggruppati in 4 macro-aree: *Business*, *Exchange*, *Structures*, *Concepts* (oltre all'area *Base* dove sono presenti elementi riusabili dagli altri per identificazione e *versioning*). Nell'area *Business* risiedono gli elementi quali il *Statistical Programme* (in processo statistico: indagine, uso di fonti amministrative, etc.) che identifica il processo che può essere ripetuto periodicamente (*Statistical Programme Cycle*). All'interno dello *Statistical Programme Cycle* c'è il *Business Process* che è un insieme di *Process Steps*. Nell'area *Concepts* vi sono tutti gli oggetti informativi di tipo metadati strutturali (concetti, popolazione, variabili, classificazioni). *Exchange* e *Structures* identificano rispettivamente i fornitori dell'informazione e i protocolli di scambio il primo e le strutture dei dati e dei metadati il secondo.

Figura 3.1 – Rappresentazione semplificata degli oggetti informativi

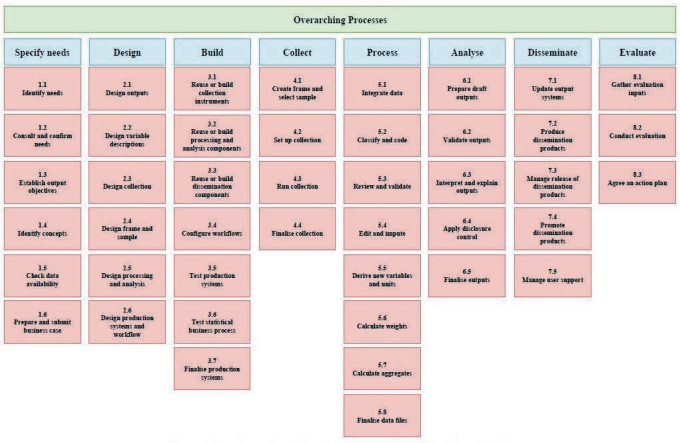


Fonte: UNECE

GSBPM descrive i processi produttivi statistici mediante uno standard di riferimento e una terminologia armonizzata. Esso si sviluppa su otto fasi, dalla definizione dei bisogni informativi, al disegno e la costruzione degli strumenti, alla raccolta dati, trattamento, analisi e diffusione. L'ultima fase include la raccolta delle evidenze e la valutazione. All'interno di ciascuna fase sono identificati dei sottoprocessi. È importante sottolineare che, nonostante graficamente il modello appaia con fasi e sottoprocessi sequenziali, la sua implementazione non è lineare, potendo alcuni sottoprocessi essere svolti prima o contestualmente ad altri che sono rappresentati successivamente.

Il modello ha anche dei processi trasversali, chiamati *Overarching Processes*, che riguardano per esempio, la gestione della qualità, l'archiviazione.

Figura 3.2 – Modello *GSBPM*

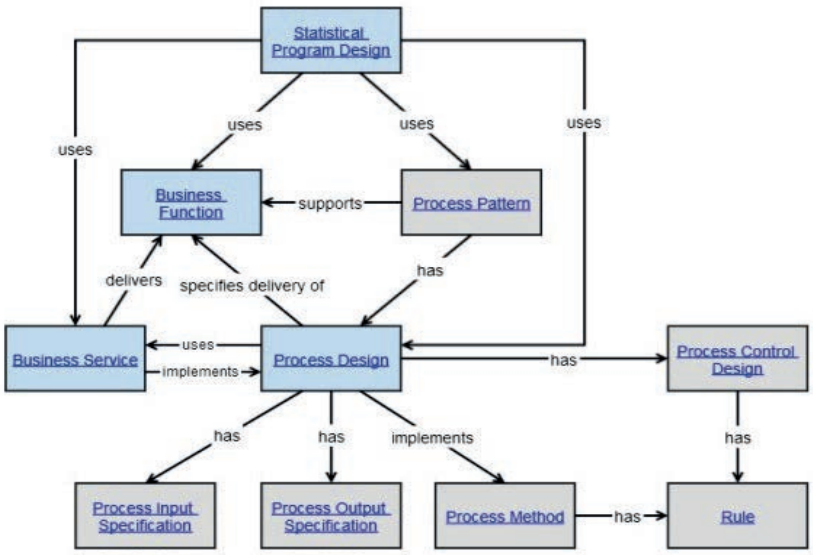


Fonte: UNECE

Allo scopo di raccogliere le informazioni utili alla fase di calcolo, si è proceduto a selezionare le sole dichiarazioni afferenti agli eventi di cui sopra. Pertanto, delle 19.995 dichiarazioni totali, solo 5.232 afferiscono alla funzione maternità e/o a congedi parentali.

La figura 3.3 che segue rappresenta sinteticamente gli elementi del modello *GSIM*, rilevanti per le finalità della modellazione dei metadati e indicatori del registro o del sistema dei registri. Questi saranno poi approfonditi e contestualizzati nel successivo sottoparagrafo.

Figura 3.3 – *Design Business Step*

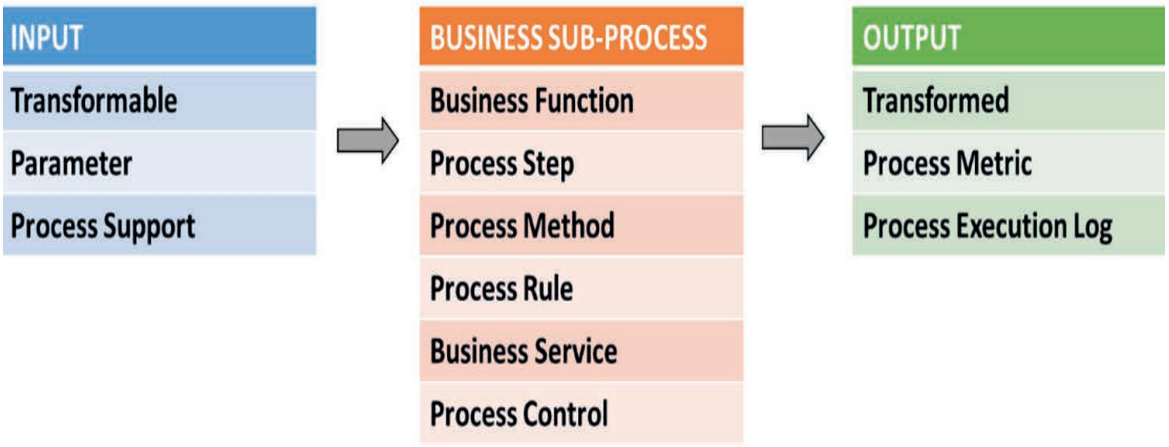


Fonte: UNECE

Il principale elemento di raccordo tra *GSIM* e *GSBPM* è rappresentato dal *Business Process*.

Una fase o un sottoprocesso di *GSBPM* (blocco centrale della figura 3.4) può essere descritto da: una funzione, dei passi, un metodo, una regola, un servizio e un *process control*. Inoltre riceve in *input* informazioni (dati trasformabili, parametri e informazioni di supporto non trasformabili) che trasforma producendo informazione in *output* (dati trasformati, metriche del processo, log di esecuzione). Gli oggetti della modellazione sono schematizzati nella figura che segue.

Figura 3.4 – Oggetti della modellazione



Fonte: Schema prodotto dagli autori

Come accennato, l'elenco di partenza degli indicatori di qualità considerati è tratto dagli indicatori standard di *performance* e qualità definiti, per ogni dimensione della qualità del prodotto, a livello Eurostat, e qui di seguito elencati brevemente.

Prospetto 3.1 – *Eurostat Quality and Performance Indicators*

Dimensione della qualità	Indicatori
Relevance	R1. *Data completeness - rate*
Accuracy and reliability	A1. *Sampling error - indicators* A2. *Over-coverage - rate* A3. *Common units - proportion* A4. *Unit non-response - rate* A5. *Item non-response - rate* A6. *Data revision - average size* A7. *Imputation - rate*
Timeliness and Punctuality	TP1. *Time lag - first results* TP2. *Time lag - final results* TP3. *Punctuality - delivery and publication*
Coerence and Comparability	CC1. *Asymmetry for mirror flows statistics - coefficient* CC2. *Length of comparable time series*
Accessibility and Clarity	AC1. *Data tables – consultations* AC2. *Metadata - consultations* AC3. *Metadata completeness - rate*

Fonte: Eurostat

4. Il contesto Istat

Le innovazioni introdotte a partire dal 2016 in Istat per modernizzare la produzione della statistica ufficiale sono state accompagnate anche da una riorganizzazione dei processi e del trattamento dei dati, secondo un modello organizzativo di *business architecture* (Istat, 2016*c*). Un aspetto molto importante di tale processo di rinnovamento dell'Istat è la centralizzazione della raccolta dati in una Direzione Centrale dedicata, con l'obiettivo di rendere più efficiente la gestione dell'*input* del sistema dei registri nonché di garantirne la valutazione della qualità.

Il disegno del sistema di supporto al monitoraggio e alla valutazione della qualità del SIR deve seguire il ciclo di vita del dato, a partire dal flusso di attività preliminari sulle fonti che entrano come *input* nella produzione e gestione dei registri che lo compongono. Inoltre, deve essere in relazione con i sistemi di monitoraggio e documentazione già esistenti all'Istat, al fine di assicurare coerenza ed evitare ridondanza.

Infatti, nonostante le principali fonti di alimentazione dei registri siano i dati di natura amministrativa, i registri sono alimentati da più tipi di *input*, riconducibili alle seguenti tipologie di dati:

- dati da fonti amministrative;
- dati da indagini Istat;
- dati da prodotti intermedi Istat;
- dati da altri Registri statistici del SIR.

Per ogni processo statistico è quindi importante contestualizzarne gli obiettivi e i collegamenti con gli altri processi, per tipologia di dati utilizzati.

Di seguito, quindi viene mostrato il sistema di pseudonimizzazione a cui i dati amministrativi vengono sottoposti, nel rispetto delle leggi preposte, e i vari sistemi di documentazione della qualità già strutturati e implementati in Istat, a cui ogni sviluppo ulteriore potrà essere agganciato coerentemente, a seconda dell'oggetto di studio.

4.1 Il processo di pseudonimizzazione dei dati amministrativi e dei dati di indagine acquisiti

Per garantire che i dati possano integrarsi correttamente tra loro nel rispetto della normativa di protezione dei dati personali (Regolamento UE 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016), a livello centralizzato viene effettuato uno specifico trattamento. Per quanto riguarda tutti gli archivi amministrativi acquisiti in Istat e utilizzati dai Registri statistici, il Sistema Integrato di Microdati (SIM) è preposto alla conservazione dei microdati amministrativi per le finalità statistiche, con particolare riguardo al rispetto delle disposizioni contenute nell'art. 25 del Regolamento 2016/679, che prevede che il titolare del trattamento, l'Istat, metta in atto misure tecniche e organizzative predefinite, *by design* e *by default* (Istat, 2021), volte

a garantire i principi di protezione dei dati[9]. In particolare, si realizzano degli specifici pretrattamenti seguendo principalmente i seguenti passi:

1. Analisi Entità/Relazioni (per gli Archivi di prima acquisizione o che hanno subito modifiche nella struttura dei dati): permette di individuare le entità (oggetti o eventi) presenti nel *dataset* e associare gli opportuni attributi per l'organizzazione del *database* relazionale SIM.
2. Analisi dei metadati (per gli Archivi di prima acquisizione o che hanno subìto modifiche nella struttura dei dati): permette la gestione dei metadati amministrativi con la definizione delle tipologie di unità statistiche incluse, delle variabili e delle classificazioni presenti.
3. Controllo dei dati e dei metadati (controlli tecnici o *technical checks*): si verifica se i dati forniti sono conformi ai dati richiesti.
4. Caricamento dei dati e dei metadati nelle opportune tabelle del *database* SIM: per gli Archivi di prima acquisizione o che hanno subìto modifiche nella struttura dei dati, vengono progettate e definite le Tabelle del *database*, scritte le procedure di caricamento e caricati i dati; per le acquisizioni consolidate, le tabelle del *database* SIM vengono alimentate con il lancio delle procedure standardizzate.

Quando negli archivi sono presenti le variabili identificative degli interessati (individui e imprese)[10], vengono effettuate anche le seguenti operazioni:

5. 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.
6. Riconoscimento delle unità statistiche di base (Individuo o Impresa) attraverso procedure di *record linkage* di tipo deterministico e apposizione del codice pseudonimo, per gli individui e per le unità economiche. Questa fase comprende, quindi, la pseudonimizzazione dei dati: per ogni nuovo *dataset* acquisito, in base alle variabili identificative disponibili e alla loro qualità, si definisce 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[11] viene abbinata a una già presente, si associa lo stesso codice pseudonimo; se la rappresentazione non si abbina, l'unità è un nuovo ingresso nel sistema, viene assegnato un nuovo codice (Runci *et al*., 2016).

Per alcuni archivi che contengono indirizzi riferiti alle unità statistiche di base, ovvero Individui o Imprese, si effettua una procedura utile all'apposizione del Codice Unico dell'Indirizzo (CUI) effettuata nell'ambito del Registro statistico base dei luoghi (RSBL).

9 In generale, l'Istat è attento alla protezione dei dati personali e si adopera continuamente per adeguare le misure all'evoluzione normativa in tutto il processo di produzione, dalle prime fasi di progettazione, secondo il paradigma della *privacy by design* e per impostazione predefinita, in conformità al paradigma della *privacy by default*.

10 Per quanto riguarda gli Individui, le variabili identificative che possono essere contenute negli Archivi amministrativi sono un sottoinsieme delle seguenti: 1) cognome; 2) nome; 3) sesso; 4) data di nascita; 5) luogo di nascita; 6) cittadinanza; 7) codice fiscale; 8) luoghi, quali indirizzo di residenza o di domicilio fiscale; 9) relazioni con altri individui di coabitazione, parentela o fiscalmente a carico. Per le imprese: 1) codice fiscale; 2) partita IVA; 3) ragione sociale.

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

Questa consiste nei seguenti passi:

7. Identificazione di tutte le rappresentazioni degli indirizzi presenti nel *dataset*.
8. Apposizione del CUI per le rappresentazioni identiche già codificate.
9. Rilascio a RSBL della lista delle rappresentazioni da codificare e acquisizione delle rappresentazioni codificate.

Per tutti gli Archivi trattati in SIM:

10. Creazione di viste sui dati per gli utenti contenenti i dati pseudonimizzati e codificati relativi al periodo di riferimento della fornitura.
11. Gestione degli accessi per gli utenti autorizzati in base alle delibere per il trattamento dei dati.

Il codice SIM per l'identificazione delle unità *target* è presente anche per i prodotti intermedi di dati, come ad esempio la Base Integrata su Istruzione e Titoli di studio (Runci *et al*., 2017), che utilizza come *input* i dati amministrativi pseudonimizzati.

Nel caso in cui i dati da indagine debbano essere integrati nei registri, analogamente a quanto viene fatto sui dati di fonte amministrativa, si procede alla pseudonimizzazione delle unità con l'apposizione del codice SIM; tale attività è gestita dal sistema SIRIL (Sistema Integrato dei dati delle RILevazioni). In genere quando le indagini utilizzano i dati amministrativi per la definizione della lista di estrazione del campione, le unità già presenti negli archivi amministrativi entrano nel campionamento già con il codice SIM e la fase di pseudonimizzazione effettuata da SIRIL coinvolge solo le nuove unità appartenenti alla popolazione *target*, ma non presenti nella lista.

In sintesi, le procedure effettuate centralmente perseguono il principio di minimizzazione di accesso ai dati personali secondo la normativa vigente e permettono di ottimizzare il processo di integrazione tra dati. 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.

Il sistema descritto è in corso di evoluzione. L'introduzione di codici pseudonimi plurimi, secondo un approccio gerarchico per domini specifici di integrazione, permetterà nel prossimo futuro di migliorare ulteriormente la gestione degli accessi ai dati seguendo i rilievi espressi dal Garante per la protezione dei dati personali.

È importante sottolineare che la procedura presentata in questo sottoparagrafo si configura come il primo sottoprocesso volto a migliorare la qualità dei dati. Infatti, i dati amministrativi si presentano spesso in formato non idoneo al loro utilizzo diretto in analisi statistiche. La procedura presentata, quindi, permette di identificare univocamente l'unità statistica di base (individuo, impresa, luogo) e di strutturare l'informazione in modo idoneo alla produzione statistica.

4.2 La documentazione della qualità

Nel corso degli ultimi decenni, l'Istat si è già dotato di un insieme di strumenti generalizzati e centralizzati per la documentazione dei metadati e della qualità.

I metadati che descrivono i dati amministrativi acquisiti, e che quindi consentono un loro corretto utilizzo per le finalità statistiche, sono documentati nella *Quality Report Card* per i dati Amministrativi (QRCA).

L'armonizzazione e la documentazione dei cosiddetti metadati strutturali per tutti i processi produttivi statistici, siano essi indagini dirette o processi che utilizzano dati di fonte amministrativa, sono assicurate all'Istat da un sistema denominato Sistema Unitario dei Metadati (SUM). Per metadati strutturali si intendono i concetti di popolazioni e unità, variabili, indicatori statistici, eccetera.

I metadati sul processo, per esempio la tecnica di raccolta dei dati per le indagini dirette, o il tipo di procedura di controllo e correzione per un qualsiasi processo produttivo statistico dell'Istat (noti come metadati referenziali) sono documentati nel Sistema Informativo sulla Qualità (SIDI-SIQual) (Brancato *et al*., 2004; Brancato *et al*., 2006).

Per quanto riguarda gli indicatori di qualità, questi sono documentati in Istat seguendo il ciclo di vita del dato statistico, dalla sua acquisizione, al trattamento dei dati, alla diffusione. In particolare, per i dati di *input* di fonte amministrativa, gli indicatori sono documentati nella già citata QRCA (Cerroni *et al*., 2014). Gli indicatori di acquisizione per le indagini dirette, gli indicatori di processo sia per le rilevazioni dirette sia per i processi che utilizzano dati di fonte amministrativa e gli indicatori di qualità dell'*output* sono standardizzati, raccolti e archiviati in SIDI-SIQual.

In particolare, la QRCA fornisce, in tempo reale, le informazioni necessarie ai processi di produzione per valutare l'usabilità dei dati amministrativi in termini di contenuti, qualità e monitoraggio delle fasi di acquisizione e di trattamento.

La QRCA è stata implementata in accordo con le *best practice* internazionali. La progettazione si avvia negli aspetti metodologici in occasione del progetto internazionale *BLUE-Enterprise and Trade Statistics (BLUE-ETS)* (Daas and Ossen, 2011) a cui l'Istat partecipa nel periodo 2011-2013, e lo sviluppo entra a regime a novembre del 2018. Il progetto *BLUE-ETS* rispondeva all'obiettivo più ampio di supportare la strategia Europea 2020 di miglioramento di un sistema di informazioni statistiche robuste e di qualità sulle imprese e il commercio.

Selezionando l'archivio di interesse si accede a informazioni relative alle iperdimensioni di fonte, metadati e dati. In particolare:

a. caratteristiche dell'archivio (denominazione standardizzata, ente titolare, trattamento previsto, rilevanza in termini di estensione di uso in Istat misurata dai lavori inclusi nel Psn che dichiarano di utilizzare l'archivio, tipi di unità, variabili e classificazioni presenti);
b. caratteristiche delle forniture (denominazione, periodicità di acquisizione, data di acquisizione, canale di acquisizione);
c. monitoraggio delle acquisizioni e dei trattamenti previsti;
d. indicatori di qualità dei dati e dei trattamenti: puntualità, tempestività, conformità (controlli tecnici dei dati amministrativi per verificarne la conformità, come analisi in serie storica del numero di *record* per ciascun file acquisito, percentuale di dati mancanti anche in serie storica, distribuzioni di frequenze dei dati categoriali, anche

in serie storica; analisi delle decodifiche mancanti), qualità delle variabili di *linkage* (variabili disponibili e numero di *record* con valori *missing* per ogni variabile), monitoraggio del processo di integrazione.

Per informazioni di dettaglio sulla QRCA si veda il volume a cura di G. Di Bella, 2021.

La QRCA si alimenta in modo automatico dai sistemi di gestione dei dati amministrativi: SIM, ARCAM[12] e dal *database* del Psn, il portale è gestito da un'interfaccia *Java* e dall'applicativo di *Business Intelligence Microstrategy*.

Per le finalità di questo lavoro, sono soprattutto gli indicatori relativi all'iperdimensione dei dati che assumono rilevanza come informazioni di *input* nella valutazione delle fasi di trattamento. Pertanto, di seguito si riporta esclusivamente questo sottoinsieme di indicatori:

- Indicatore di puntualità della fornitura dell'archivio amministrativo;
- Indicatore di tempestività della fornitura da parte dell'Ente e di tempestività complessiva;
- Serie storica del numero di *record* per ciascun file acquisito nella fornitura;
- Serie storica dei dati mancanti per le principali variabili dei *dataset*;
- Frequenze delle distribuzioni dei dati categoriali, anche in serie storica;
- Decodifiche mancanti;
- Indicatori di qualità del *linkage* per l'abbinamento per l'attribuzione del codice SIM.

Attualmente la QRCA si avvale anche dell'integrazione con SIDI per il calcolo degli indicatori di rilevanza, sfruttando l'associazione tra lavoro Psn e normativa europea presente in SIDI-SIQual. Il collegamento è, quindi, a livello di lavoro Psn. In futuro si progetta di effettuare il collegamento anche a livello di archivi amministrativi al fine di rendere la documentazione dei processi di SIDI-SIQual coerente alla QRCA rispetto agli archivi utilizzati come *input* e creare una maggiore sinergia.

La futura integrazione tra QRCA e SUM sarebbe utile per l'implementazione degli indicatori di comparabilità a livello concettuale tra variabili amministrative (*input* dei registri) e variabili statistiche (*output* dei registri).

Quando il processo produttivo statistico è basato su un disegno di indagine, ossia vi è una lista di riferimento per l'estrazione del campione oppure l'indagine è censuaria e i dati sono raccolti mediante un questionario, sia esso elettronico o meno, gli indicatori di qualità sui dati raccolti sono documentati e archiviati nel sistema SIDI-SIQual, insieme ai metadati rilevanti. Si tratta di indicatori di qualità sulla lista di riferimento (copertura, errori di lista) e di indicatori sull'esito della fase di intervista, noti come indicatori di mancata risposta totale con i motivi della non risposta.

In base alle fasi effettuate nell'ambito dello specifico processo produttivo statistico in SIDI-SIQual sono calcolati anche indicatori sul sottoprocesso di codifica (ossia di

12 Arcam https://arcam.istat.it : Sistema di acquisizione di archivi amministrativi e dati di indagine. Si tratta di un applicativo gestionale per l'acquisizione degli archivi il cui accesso è consentito al personale del settore della raccolta dei dati.

trasformazione delle risposte testuali in codici) e di controllo e correzione dei dati. Infine il sistema documenta anche indicatori sulla qualità del dato statistico pubblicato, per esempio la sua tempestività, le variazioni legate alla politica di revisione per statistiche di natura congiunturale, la coerenza delle stime con quelle provenienti da altri processi relativi allo stesso dominio e la loro comparabilità nel tempo.

Schematizzando, i gruppi di indicatori gestiti in SIDI-SIQual sono:

Indicatori orientati alla qualità del processo:

- Copertura;
- Mancata risposta totale;
- Codifica;
- Controllo e correzione;
- Costi.

Indicatori orientati alla qualità del prodotto:

- Analisi delle revisioni;
- Tempestività e puntualità;
- Coerenza tra stime provvisorie e definitive;
- Coerenza con altre fonti;
- Lunghezza della serie storica confrontabile.

Per ogni gruppo vi sono poi indicatori di dettaglio: ad esempio, l'indicatore di copertura consta di indicatori di dettaglio relativi alla sovracopertura, ulteriormente distinti per i motivi (unità non più esistente, cambiamento di strato, unità fuori *target*). Alcuni indicatori sono armonizzati e coincidenti con quelli adottati nel Sistema Statistico Europeo e già presentati nel Prospetto 3.1.

SIDI-SIQual, oltre agli indicatori di qualità, documenta i metadati referenziali sui processi statistici ed è integrato con diversi sistemi informativi dell'Istat, al fine di ridurre il più possibile il *burden* di documentazione e di mantenere coerenza nelle informazioni presenti nei vari sistemi. In particolare SIDI è collegato al Psn e mantiene l'associazione tra i lavori Psn e i processi statistici dell'Istat. SIDI-SIQual è inoltre integrato con SUM rispetto ad alcune aree informative: mette infatti a disposizione di SUM alcune informazioni, quali i processi, le unità statistiche e i questionari. Sono in corso degli sviluppi per rafforzare l'integrazione rispetto alle unità statistiche e far sì che siano allineate nei due sistemi, ed estenderla alle aree delle classificazioni e delle variabili. I metadati e gli indicatori in SUM, QRCA e SIDI-SIQual offrono dei contribuiti per singole parti del ciclo di vita del dato statistico e in particolare per specifiche tipologie di processi statistici di produzione. Alcuni di essi possono essere estesi anche all'ambito del sistema dei registri SIR, tuttavia è importante costruire un sistema specifico e di supporto ai registri, interoperabile o integrato con i sistemi esistenti. Nel lungo periodo è inoltre previsto di progettare e sviluppare un nuovo sistema di metadati unico per sostituire sia SIDI-SIQual sia SUM, anche di questo andrà tenuto conto nello sviluppo del sistema di supporto ai registri.

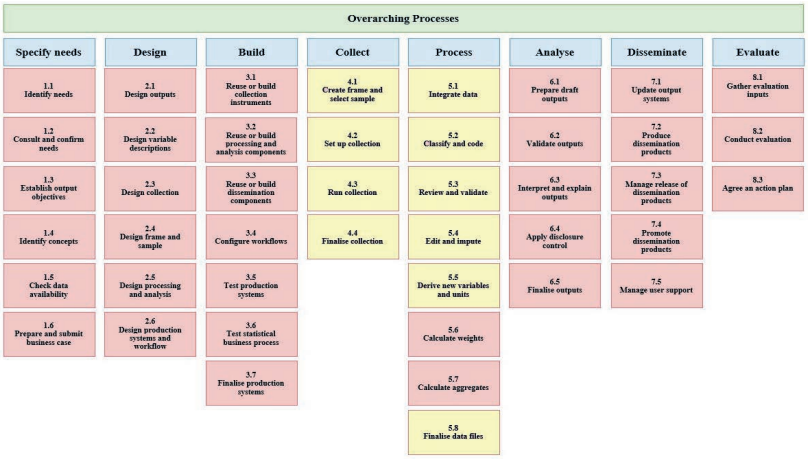
5. Modellazione concettuale per i metadati e gli indicatori di qualità per il Sistema dei Registri

5.1 Contestualizzazione della proposta per il SIR in relazione agli standard internazionali

L'obiettivo del lavoro è la realizzazione di un quadro di riferimento per la documentazione dei processi di costruzione e di aggiornamento a regime di un sistema di registri, in particolare il SIR. Ovvero, si considera una situazione dove la progettazione è già stata realizzata, e l'*output* finale che deve essere documentato è il registro stesso. Pertanto, in questo lavoro non si considera la documentazione o la qualità delle stime derivate dal registro.

Ne consegue che le fasi del *GSBPM* da considerare sono quelle prettamente esecutive, cioè a dire le fasi di acquisizione dei dati e trattamento (*Collect* e *Process* in *GSBPM*) a esclusione delle sottofasi di calcolo dei pesi e degli aggregati. Nella Figura 5.1 che segue, le sottofasi da esaminare sono evidenziate in colore giallo.

Figura 5.1 – Fasi/sottoprocessi del *GSBPM* prese in considerazione



Fonte: UNECE

I sottoprocessi considerati rilevanti sono stati personalizzati per il sistema dei registri. Il modello adottato, che segue *GSIM*, prevede che tutta la documentazione e la definizione degli indicatori sia costruita intorno allo specifico sottoprocesso, o per entità più di dettaglio all'interno del sottoprocesso, denominate *process step*. Inoltre, lo schema di riferimento è stato integrato con elementi aggiuntivi per conciliare l'approccio standard con le esigenze interne all'Istat.

Nel Prospetto 5.1 sono descritti gli elementi utilizzati per la documentazione e la costruzione degli indicatori per il SIR, insieme alle definizioni e a esempi. Questi sono stati sviluppati in coerenza con il lavoro svolto a livello internazionale (UNECE, 2020), ma tenendo conto del contesto e delle esigenze specifiche dell'Istituto.

Le definizioni riportate non coincidono in modo rigoroso con quelle riportate in *GSIM* ma sono state semplificate. Inoltre, *GSIM* tende a definire oggetti speculari a livello di disegno e di esecuzione. Ciò riflette due approcci sostanzialmente differenti, rispettivamente una finalità documentativa e un effettivo supporto per l'implementazione del processo. Per chiarezza, se si documenta la formula di una regola del controllo e correzione, nel primo caso è sufficiente documentare la formula, mentre nel secondo è indispensabile avere un sistema (parametrizzato) per l'applicazione della formula ai dati. Sarà obiettivo della fase di implementazione del sistema di monitoraggio e valutazione della qualità del SIR, stabilire quale approccio perseguire.

Prospetto 5.1 – Elementi per la documentazione del processo

Macro-elemento	Nome oggetto	Definizione ed Esempio
Input	*Input trasformabile (Transformable)*	Oggetti forniti in *input* al sottoprocesso per essere da esso modificati. Es.: Nel sottoprocesso di controllo e correzione, il *dataset* (con la sua struttura) che è sottoposto a controllo e correzione.
	Parametri *(Parameter)*	Oggetti forniti in *input* al sottoprocesso per configurare il sottoprocesso stesso. Es.: I parametri di un modello di stima.
	Input ausiliario *(Process support)*	Oggetti forniti in *input* al sottoprocesso, necessari a esso, ma non sono da questo modificati. Es.: La classificazione standard usata in un sottoprocesso di codifica.
Sottoprocesso	Obiettivo *(Business Function)*	Obiettivo del sottoprocesso o attività. Es.: Ottenere un *dataset* completo e coerente; Correggere gli errori sistematici.
	Fase *(Business process GSIM)*	Concetto *GSIM* che si può mappare con la fase *GSBPM* e che per definizione è composto da una serie di sottoprocessi o attività (*process step*). Es.: Trattamento (*Process*).
	Sottoprocesso o attività *(Process step)*	Sottoprocesso (*GSBPM*) o attività più elementare che realizza un obiettivo specifico. Può a sua volta essere formato da altri sottostep. Es.: Controllo e correzione; Individuazione degli *outlier*.
	Azioni di controllo della qualità	Azioni di prevenzione, monitoraggio e valutazione a posteriori degli errori Es.: Prevenzione dell'errore di mancata risposta totale.
	Metodologia (*Process Method*)	Descrizione della metodologia usata. Es.: Nel controllo e correzione, applicazione del metodo del donatore.
	Regola (*Rule*)	Regola algebrica o matematica che si può applicare nel metodo o per gestire il passaggio da un sottoprocesso al successivo. Es.: Se $x_1 < V_1 < x_2$ allora V_1 = errore.
	Servizio statistico (*Business Service*)	L'interfaccia o un modo di svolgere la funzione (o obiettivo). Può rappresentare un contratto di servizio, non necessariamente IT. Es.: *Software Selemix* (*Selective editing via Mixture models*) per l'individuazione degli errori potenzialmente influenti.
	Workflow (*Process Control*)	Elemento che indica l'attività successiva a quello in oggetto. Può riflettere un processo decisionale. Es.: Da *editing* automatico su errori di dominio ed errori sistematici a imputazione deterministica.
Output	*Output* trasformato *(Transformed output)*	Oggetti creati o modificati dal sottoprocesso. Es.: Il *dataset* imputato o una nuova variabile creata.
	Qualità e *performance* (*Process Metric*)	Sottoprodotto del sottoprocesso, utile a documentare, misurare e valutare l'esecuzione del sottoprocesso stesso. Es.: Tasso di imputazione.
	Log della procedura (*Process Execution Log*)	Rapporto di esecuzione del sottoprocesso, essenzialmente legato all'esecuzione pratica (*log* del programma, tempo di esecuzione della procedura). I dati registrati nel *log* della procedura possono essere utilizzati per la costruzione di una *process metric*.

Fonte*:* Eurostat

In generale, i termini utilizzati in *GSIM* vengono tradotti Dove ritenuto opportuno, viene utilizzata una terminologia più intuitiva rispetto a quella di *GSIM*. In questo caso nel prospetto verrà evidenziato in parentesi il termine *GSIM*. Alcuni oggetti non sono inclusi in quanto non necessari per un sistema di documentazione e valutazione, ma utili solo per un sistema che opera effettivamente sui dati. L'elemento che identifica la gestione del flusso viene considerato sia all'interno del sottoprocesso sia tra sottoprocessi (macroelemento: Relazione tra sottoprocessi).

L'implementazione sarà graduale e, in una prima fase, potrà non coinvolgere tutti gli elementi identificati.

5.2 La documentazione per il SIR

Nel predisporre la documentazione dei processi e della qualità del sistema dei registri, è necessario tener conto di varie esigenze e dell'articolazione del sistema stesso. Oltre a considerare l'usuale dicotomia tra documentazione e qualità del processo e del prodotto, sarà utile includere anche delle ulteriori categorie di documentazione.

La ***documentazione generale*** inquadra i singoli registri del SIR in termini "demografici" e di obiettivi, ed è funzionale all'identificazione dell'oggetto registro e agli aspetti di gestione di tipo anche amministrativa.

Per ***documentazione e qualità del processo*** si intende l'insieme di informazioni attinente alle operazioni di trasformazione di un *input* in un *output* e tutte le misure di qualità relative, come modellate nel precedente sottoparagrafo.

Per ***documentazione e qualità del prodotto*** si intende l'insieme dei metadati strutturali e referenziali attinenti ai microdati di *input* e di *output* (ovviamente alcuni *output* corrispondono ad *input* di una fase successiva e possono essere considerati intermedi, mentre il primo *input* e l'ultimo *output* sono univoci). La qualità di prodotto è strettamente in relazione con la qualità del processo in quanto dettaglia la qualità dell'*input* trasformabile e dell'*output* trasformato.

Infine, i registri che compongono il SIR sono in relazione tra di loro. Tale relazione è mappata attraverso il registro delle attività. Sarà quindi necessario stabilire le misure di ***qualità delle relazioni*** tra registri del sistema. Questa area sarà sviluppata in un successivo momento.

Queste aree di documentazione sono presentate nelle successive sessioni. Esse, come sarà evidente dalla trattazione, presentano diversi gradi di sviluppo e diversi livelli di maturità.

5.2.1 Documentazione generale

Per ciascun registro che compone il SIR, le informazioni generali da documentare sono riportate nel Prospetto 5.2.

Prospetto 5.2 – Documentazione generale

Scheda anagrafica del registro	Denominazione Acronimo Codice Psn Struttura Responsabile Primo anno di costruzione Tipo (Base/Esteso/Tematico) Periodicità di aggiornamento Periodicità di rilascio (consolidato) Periodicità di rilascio (versioni intermedie) Regolamenti
Obiettivi informativi del registro	Descrizione Popolazione *target* Principali variabili target con definizione
Fonti del registro	Denominazione Fonte Fornitore [Istat, Nome Ente non Istat] Tipo di Fonte [Dati da fonti amministrative, Dati da indagini Istat, Dati da prodotti intermedi Istat, Dati da altri registri statistici del SIR] Periodicità della fornitura Modalità di acquisizione Stadio di lavorazione dell'*input* [provvisorio, definitivo]
Indicatori di qualità	Rilevanza delle fonti per le finalità del registro Tempestività e puntualità delle fonti

Fonte*:* Schema prodotto dagli autori

Prospetto 5.3 – Indicatori di qualità

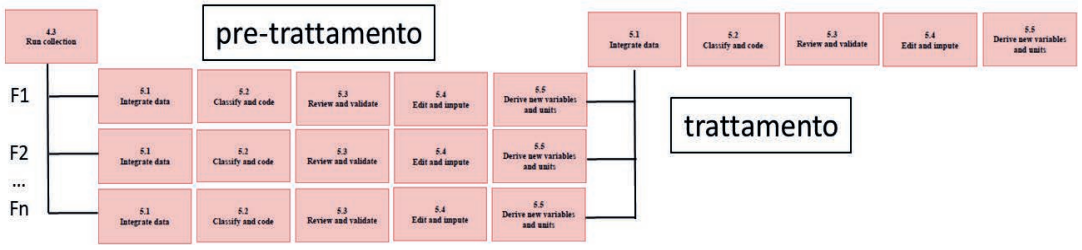
Indicatori di qualità	Rilevanza delle fonti per le finalità del registro Tempestività e puntualità delle fonti

Fonte*:* Schema prodotto dagli autori

5.2.2 Documentazione e qualità del processo

Per quanto riguarda la qualità del processo, i sottoprocessi rilevanti del *GSBPM* sono schematizzabili come mostrato nella Figura 5.2. Da notare che non tutti i sottoprocessi devono essere necessariamente presenti e svolti nella sequenza riportata in figura.

Figura 5.2 – Schematizzazione dei sottoprocessi documentati



Fonte: Schema prodotto dagli autori

Il primo sottoprocesso che si considera è *Run collection* (4.3), inteso in questo contesto come attività di acquisizione delle fonti utilizzate per il singolo registro del SIR. Le fonti possono essere di diversa tipologia come riportato nel Paragrafo 4.

Supponendo che si acquisiscano *n* fonti (F_1, F_2,… , F_n), il modello adottato organizza i sottoprocessi in:

- attività di pre-trattamento delle fonti;
- attività di trattamento delle fonti dopo la loro integrazione.

Nell'ambito del pre-trattamento delle fonti, i sottoprocessi possono essere svolti sulle singole fonti oppure già su una loro parziale integrazione. Questi sono in genere orientati a:

- controllare e migliorare la qualità delle singole fonti (calcolarne la copertura, rimuovere errori sulle unità e sulle variabili);
- riconciliare le definizioni e classificazioni e trasformare i dati in modo da armonizzarli;
- applicare delle vere e proprie procedure di controllo e correzione;
- creare nuove unità e/o variabili. Le stesse attività possono essere ripetute su tutte le fonti integrate (trattamento).

Esse sono documentabili in modo analogo a quello utilizzato per il trattamento delle singole fonti.

Queste attività sono mappabili all'interno dei sottoprocessi del *GSBPM*.

In particolare sono stati considerati i sottoprocessi elencati nel Prospetto 5.4. che sono stati sviluppati nelle schede di documentazione del sottoparagrafo successivo.

Prospetto 5.4 – Sottoprocessi sviluppati per la documentazione del SIR

#	Sottoprocesso o *step*	Possibili obiettivi
1	Acquisizione dei dati della fonte (4.1. *Run Collection*)	Acquisizione dei dati (interna da Direzione Centrale della Raccolta Dati dell'Istat).
2	Integrazione dei dati (5.1. *Integrate*)	Integrazione statistica tra le fonti (*Record linkage*, *Statistical matching*). Integrazione finalizzata: - alla deduplicazione sui singoli *dataset* acquisiti/*dataset* integrato; - alla misurazione di indicatori di copertura; - a elaborare priorità tra fonti se la stessa variabile è presente per le stesse unità in varie fonti.
3	Ricodifica (5.2. *Classify and Code*)	Ricodifica di variabili (alfanumeriche in numeriche, numeriche in classi diverse, etc.) sui singoli *dataset* acquisiti/*dataset* integrato.
4	Validazione dei microdati (5.3. *Review & validate*)	Controlli di qualità e coerenza sui singoli *dataset* acquisiti/*dataset* integrato.
5	Controllo e correzione (5.4. *Edit & Impute*)	Procedure di controllo e correzione sui singoli *dataset* acquisiti/*dataset* integrato.
6	Derivazione di nuove unità e variabili (5.5. *Derive new variables and units*)	Creazione di nuove unità, Creazione di nuove variabili.

Fonte: Schema prodotto dagli autori

Da sottolineare come il sottoprocesso di integrazione abbia un'accezione ampia e includa, oltre alle operazioni di integrazione tra *dataset* (*record linkage*, *statistical matching*), anche un'attività interna a singoli *dataset*, ossia il trattamento dei dati per rimuovere le duplicazioni nonché l'integrazione finalizzata alla stima della copertura.

Le schede di modellazione del processo e degli indicatori sono nel sottoparagrafo 5.3.

5.2.3 Documentazione e qualità del prodotto

Ciascuna delle schede documentative precedentemente descritte include degli *input* e *output*.

La qualità del prodotto in *input* al sottoprocesso di acquisizione dei dati è già disponibile attraverso la QRCA o SIDI-SIQual rispettivamente per dati di fonte amministrativa e dati di indagine. In questo caso si rinvia a tali appositi sistemi (si veda sottoparagrafo 3.1), ma questo non avviene per gli *input-output* degli altri sottoprocessi. È quindi necessario contemplare nel sistema di documentazione del SIR anche la documentazione in termini di metadati e qualità degli altri prodotti di *input* e dell'*output* finale. Non viene incluso a questo stadio dei lavori l'*output* di tipo macro, ossia le stime prodotte dai dati.

Prospetto 5.5 – Qualità dell'*input* /*output* di tipo micro

Input trasformabile o *Output* trasformato	Metadati strutturali, Popolazione, Tipo di unità [individui, imprese, istituzioni, famiglie], Variabili [nome, definizione] e *codelist* e classificazione, Periodo di riferimento dei dati, Indicatori di qualità, Sulle variabili: Tasso di completezza delle variabili, Coerenza con altre fonti, Misure di qualità, Stima della varianza e della distorsione dovuta all'errore di misurazione.

Fonte*:* Schema prodotto dagli autori

5.3 Schede di modellazione per la documentazione dei processi del SIR

In questo sottoparagrafo si riportano le schede sviluppate per rappresentare e documentare il processo di un registro a regime e la sua qualità (Prospetti 5.6 - 5.11). Le schede si ispirano al *GSBPM*, con l'obiettivo di rappresentarne i sottoprocessi e sono personalizzate per riflettere il contesto produttivo statistico dell'Istat. Poiché l'obiettivo principale è quello della valutazione della qualità, le schede includono anche gli indicatori di qualità identificati e mappati nella struttura dei sottoprocessi del *GSBPM*.

È opportuno sottolineare nuovamente che, come nel *GSBPM*, l'ordine con cui i sottoprocessi sono applicati – e quindi delle schede che li descrivono – non è prestabilito; inoltre un sottoprocesso può essere utilizzato più volte e alcuni sottoprocessi possono essere diversi per variabili diverse. Quindi, ci si aspetta che per ogni processo, la rappresentazione del processo attraverso le schede possa essere molto vario, determinando sequenze anche diverse tra loro.

È altresì importante notare che le schede possono riguardare un singolo *dataset* prima della sua integrazione, oppure un *dataset* frutto dell'integrazione di due o più fonti. Ciò sarà evidente dalla documentazione specifica fornita per l'*input*.

L'insieme degli indicatori di qualità proposti è ampio e gli indicatori non sono tutti obbligatori ma anzi vanno selezionati e interpretati nel contesto specifico del processo sotto studio. Per esempio, se si stanno integrando due fonti che dovrebbero contenere le stesse unità, allora gli indicatori di integrazione possono essere di supporto alla valutazione della qualità mentre, se le due fonti riguardano popolazioni sostanzialmente diverse e complementari allora, ovviamente non ci si attende una elevata sovrapposizione tra di esse.

Gli indicatori proposti sono in generale autoesplicativi, tuttavia nell'Appendice vengono fornite le formule di calcolo e un supporto alla loro interpretazione.

Prospetto 5.6 – Sottoprocesso 1: Acquisizione dei dati della fonte

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset* (dati di fonte amministrativa, dati di indagine, dati di altro registro statistico) (unità e variabili).
	Parametri	
	Input ausiliario	Documentazione della QRCA.
Sottoprocesso	Obiettivo	Acquisizione dei dati.
	Fase	Raccolta (4. *Collect in GSBPM*).
	Sottoprocesso o attività	Acquisizione dei dati (4.4. *Run collection in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di caricamento dei dati della fonte.
	Metodologia	
	Regola	Accesso ai dati condizionato alle delibere di autorizzazione al trattamento di dati personali.
	Servizio statistico	Abilitazione all'accesso ai dati tramite utenza personale.
	Workflow	
Output	*Output* trasformato	*Dataset* acquisito (dati di fonte amministrativa, dati di indagine, dati di altro registro statistico) (unità e variabili).
	Qualità e *performance*	1.1. Numero di *record* acquisiti nella fornitura rispetto allo stesso numero nella fornitura precedente, 1.2. Documentazione delle variazioni nella struttura dei dati e nei metadati.
	Log della procedura	Numero di *record* rigettati dalla procedura di caricamento dei dati.

Fonte*:* Schema prodotto dagli autori

Prospetto 5.7 – Sottoprocesso 2: Integrazione

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset*1, *Dataset*2, …(unità e variabili).
	Parametri	Funzione di confronto, Soglie, Chiavi di *linkage*, variabili di *blocking.*
	Input ausiliario	Codici SIM.
Sottoprocesso	Obiettivo	Integrare i dati (deduplicare, integrare unità, integrare variabili, stimare la copertura).
	Fase	Trattamento (5. *Process in GSBPM*).
	Sottoprocesso o attività	Integrazione (5.1. *Integrate in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di integrazione.
	Metodologia	*Record linkage* deterministico, deterministico gerarchico, probabilistico, *privacy preserving and predictive linkages (classification or regression techniques), statistical matching.*
	Regola	Modello di integrazione, Relazione 1 - 1, *n* - 1, *n* - *n*.
	Servizio statistico	*Relais (REcord Linkage At IStat), Package R Statmatch.*
	Workflow	Gestione del flusso sequenziale e condizionale tra metodi diversi all'interno dell'integrazione.
Output	*Output* trasformato	*Dataset* integrato, *dataset* al netto delle duplicazioni, *Dataset* dei *record* non abbinati.
	Qualità e *performance*	2.1. Percentuale di unità duplicate, 2.2. Tasso di *link* (o *match rate*), 2.3. Tasso di falsi *link*, 2.4. Tasso di falsi *non-link*, 2.5. Tasso di dati mancanti o errati nella variabile di *linkage*, 2.6. Distanze tra distribuzioni sulle variabili rilevanti nel *dataset* integrato e nei *dataset* di *input*.
	Log della procedura	Tempi di integrazione.

Fonte*:* Schema prodotto dagli autori

Prospetto 5.8 – Sottoprocesso 3: Codifica

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset* (singola fonte o integrato) (unità e variabili).
	Parametri	Nome variabile, valori da sottoporre a codifica, soglie di validità.
	Input ausiliario	Classificazione obiettivo, Dizionario, Manuale di regole.
Sottoprocesso	Obiettivo	Codifica delle risposte aperte.
	Fase	Trattamento (5. *Process in GSBPM*).
	Sottoprocesso o attività	Classifica e codifica (5.2. *Classify and Code in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di classificazione.
	Metodologia	Codifica manuale, interattiva, automatica.
	Regola	Regole di codifica manuali o basate sul modello del sistema esperto: • *dictionary algorithms*: algoritmi che si avvalgono di parole (o gruppi di parole) particolarmente informative per determinare univocamente l'assegnazione del codice; • *weighting algorithms*: ricerca di *match* esatti o parziali sulla base di funzioni di similarità tra testi dove alle parole è attribuito un peso, empirico o probabilistico, proporzionale al loro grado informativo; • *sub-strings algorithms*: ricerca di *match* basati sull'accoppiamento di bigrammi o trigrammi di testo.
	Servizio statistico	*Circe v.1. (Comprehensive Istat R Coding Environment)*.
	Workflow	Gestione del flusso sequenziale e condizionale tra metodi diversi all'interno della codifica.
Output	*Output* trasformato	*Dataset* + variabile codificata *(dataset fonte/integrato)*.
	Qualità e *performance*	3.1. Tasso di efficacia della codifica manuale, 3.2. Tasso di efficacia della codifica interattiva, 3.3. Tasso di efficacia della codifica automatica , 3.4. Indicatori di carico sui codificatori: numero di codificatori, 3.5. Indicatori di carico sui codificatori: *workload*, 3.6. Tasso di mancata codifica.
	Log della procedura	Tempi codifica.

Fonte*:* Schema prodotto dagli autori

Prospetto 5.9 – Sottoprocesso 4: Validazione dei microdati

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset* (singola fonte o integrato) (unità e variabili).
	Parametri	Variabili che devono essere validate, Assunzioni delle regole, Valori soglia.
	Input ausiliario	*Dataset* ausiliario.
Sottoprocesso	Obiettivo	Ottenere un *dataset* validato.
	Fase	Trattamento (5. *Process in GSBPM).*
	Sottoprocesso o attività	Revisione e validazione dei microdati (5.3. *Review and validate in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di *range* e di coerenza.
	Metodologia	
	Regola	Regole di individuazione di errori e incoerenze.
	Servizio statistico	
	Workflow	Gestione del flusso sequenziale e condizionale tra metodi diversi all'interno della procedura di controllo e correzione.
Output	*Output* trasformato	*Dataset* validato.
	Qualità e *performance*	4.1. Percentuale di unità presenti più volte con gli stessi identificativi all'interno della stessa fonte, 4.2. Percentuale di unità eliminate perché non utilizzabili, 4.3. Numero di unità con almeno un valore mancante, 4.4. Tasso di mancata risposta parziale, 4.5. Numero di variabili che falliscono almeno una regola di *edit* per tipo di regola (*soft/hard*), 4.6. Percentuale unità con almeno una regola violata (*edit failure rate*).
	Log della procedura	Tempi di esecuzione della procedura, eventuali errori nell'esecuzione della procedura.

Fonte*:* Schema prodotto dagli autori

Prospetto 5.10 – Sottoprocesso 5: Controllo e correzione (a)

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset* (singola fonte o integrato) (unità e variabili).
	Parametri	Variabili che devono essere controllate e corrette, Assunzioni del modello (ad esempio normalità della distribuzione, etc.), Parametri del modello di controllo e correzione, Valori soglia.
	Input ausiliario	*Dataset* ausiliario.
Sottoprocesso	Obiettivo	Ottenere un *dataset* completo e coerente.
	Fase	Trattamento (5. *Process in GSBPM*).
	Sottoprocesso o attività	Controllo e correzione (5.4. *Edit and impute in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di misurazione e mancata risposta parziale.
	Metodologia	Metodi di revisione (b): *Edit rules,* *Analytical methods for review,* *Sufficiency study of unit,* *Micro-level consistency,* *Score by auxiliary variable,* *Score calculation for selective editing,* *Interactive review of unit.* Metodi di selezione: *Selection by scores,* *Selection by structure,* *Macro-level selection,* *Interactive unit selection,* *Micro-level selection of variables,* *Macro-level selection of variables,* *Interactive variable selection.* Metodi di trattamento: *Interactive treatment of errors,* *Deductive imputation,* *Model based imputation,* *Donor imputation,* *Consistency adjustment,* *Unit rejection,* *Unit creation,* *Unit linkage.* (fonte: *GSDEM*)
	Regola	Regole di individuazione degli errori e modelli di trattamento.
	Servizio statistico	*Banff*, *Canceis* (*CANadian Census Edit and Imputation System*), ConcordJava (CONtrollo e CORrezione dei Dati versione con interfaccia *Java*) e SeleMix (*Selective editing via Mixture models*).
	Workflow	Gestione del flusso sequenziale e condizionale tra metodi diversi all'interno della procedura di controllo e correzione.
Output	*Output* trasformato	*Dataset* editato/imputato/validato.
	Qualità e *performance*	Misure di qualità relative ai dati grezzi: 5.1. Numero di unità con almeno un valore mancante, 5.2. Numero di unità che falliscono almeno una regola di *edit* (*edit failure rate*), 5.3. Tasso di mancata risposta parziale. Misure di qualità relative alla procedura di imputazione: 5.4. Tasso di imputazione, 5.5. Tasso di imputazione per variabile, 5.6. Tasso di modificazione per variabile, 5.7. Tasso di imputazione netta per variabile, 5.8. Tasso di cancellazione per variabile, 5.9. Tasso di imputazione ponderato . (Per un insieme più dettagliato di possibili indicatori si può fare riferimento a EDIMBUS)
	Log della procedura	Tempi di esecuzione della procedura, eventuali errori nell'esecuzione della procedura.

Fonte*:* Schema prodotto dagli autori

(a) Andrà valutato se dividerlo in due *step* (uno controllo e l'altro correzione) o rifarsi alla divisione nelle 3 macro funzioni di *GSDEM*: Revisione/Selezione/Trattamento.

(b) Poiché per questo sottoprocesso c'è il riferimento internazionale di *GSDEM* si è fatto riferimento a esso per l'elenco dei metodi. Questi devono essere verificati per selezionare quelli maggiormente adottati all'Istat.

Prospetto 5.11 – Sottoprocesso 6: Derivazione di nuove unità e variabili

Macro-elemento	Nome oggetto	Elemento o valore
Input	*Input* trasformabile	*Dataset* (singola fonte o integrato) (unità e variabili).
	Parametri	Variabili e unità, funzione di trasformazione/derivazione.
	Input ausiliario	*Dataset* amministrativo.
Sottoprocesso	Obiettivo	Ampliare, armonizzare o standardizzare l'informazione statistica.
	Fase	Trattamento (5. *Process in GSBPM*).
	Sottoprocesso o attività	Derivazione unità e variabili (5.5. *Derive new variable and units in GSBPM*).
	Azioni di controllo della qualità	Prevenzione, monitoraggio e valutazione degli errori di allineamento e identificazione delle unità e di misurazione delle nuove variabili.
	Metodologia	Funzioni di trasformazione delle unità (ricostruzione semplice per esempio per aggregazione, assistita da esperto, assistita da integrazione con dati, mista); Funzione di trasformazione delle variabili deterministica, modello casuale.
	Regola	Regole di derivazione secondo formule matematiche o assunzioni da modello (per esempio: aggregazione, soppressione, arrotondamento). Ordine di derivazione (alcune variabili derivate possono essere a loro volta basate su variabili derivate).
	Servizio statistico	PROCEDURA FAMIGLIE per l'individuazione di tipologie familiari.
	Workflow	Gestione del flusso sequenziale tra differenti funzioni di trasformazione dall'interno della procedura di derivazione.
Output	*Output* trasformato	*Dataset* con nuove variabili e unità.
	Qualità e *performance*	6.1. Numero di unità derivate o trasformate, 6.2. Tasso di copertura delle unità derivate o trasformate, 6.3. Indicatori di confronto su variabili tra unità derivate e non derivate.
	Log della procedura	Tempi di esecuzione della procedura, eventuali errori nell'esecuzione della procedura.

Fonte*:* Schema prodotto dagli autori

5.4 Applicazione al Registro Esteso Frame-SBS

In questo sottoparagrafo si presenta un primo esercizio di schematizzazione, secondo il modello proposto, di un registro analizzato durante il gruppo di lavoro, il cui grado di consolidamento permette un'analisi strutturata.

Il registro Frame-SBS presenta già uno schema di validazione e monitoraggio proprio, sviluppato secondo criteri coerenti con gli schemi presentati, nel rispetto della tradizione Istat in tale campo e illustrato in alcuni lavori dove sono presentate le misurazioni degli indicatori proposti (Curatolo *et al.*, 2016; Luzi *et al.*, 2017). Rappresenta, quindi, un ottimo caso di studio per testare come le innovazioni possono essere recepite dai nuovi processi e implementate in modo efficiente.

Questo primo esercizio permette quindi di testare l'applicabilità del modello teorico di rappresentazione proposto e porre le basi conoscitive per evidenziare se eventualmente ci siano dei punti da approfondire, in termini di misurazione della qualità rispetto alle varie fasi del processo secondo il modello di metadatazione proposto.

5.4.1 Schematizzazione del processo e della misurazione della qualità di Frame – SBS

In questo sottoparagrafo si presenta una prima rappresentazione, secondo la modellazione proposta, del processo del registro esteso Frame - SBS.

Fase 1. Acquisizione delle fonti.

Nella descrizione della produzione o aggiornamento a regime del registro, partiamo dalla fase di acquisizione delle fonti che in GSBPM rientra nel sottoprocesso:

4.3 *Run Collection* (Acquisizione interna).

Il Registro acquisisce dalla Direzione Centrale della Raccolta Dati dell'Istat le fonti della seguente tipologia:

- Archivi amministrativi (Bilanci da Infocamere, Modelli vari da Agenzie delle Entrate);
- Registri Statistici dell'Istat: Registro annuale su retribuzioni, ore e costo del lavoro individuale (RACLI[13]), Registro Statistico delle Imprese Attive inclusa la parte relativa ai dati sull'occupazione (ASIA e ASIA occupazione).

Tali fonti hanno ruoli diversi. ASIA identifica la popolazione obiettivo.

L'acquisizione avviene, previa notifica da parte della Direzione Centrale della Raccolta Dati dell'Istat, attraverso il *download* delle sotto-tabelle di interesse.

Le fonti amministrative (anno di riferimento dei dati *t*) sono disponibili a gennaio dell'anno $t + 2$, a eccezione dei Bilanci civilistici che sono disponibili a marzo. Anche le fonti statistiche (ASIA, ASIA Occupazione e RACLI) sono messe a disposizione a marzo dello stesso anno. I dati sono corredati dal codice SIM assegnato alle unità economiche.

Fase 2. Pre-trattamento delle fonti.

Tutte le fasi di pre-trattamento delle fonti vengono effettuate su tutte le unità con codice SIM.

Il pre-trattamento consiste in tre principali passi del processo: l'identificazione e l'eliminazione di duplicazioni di unità, l'eliminazione di unità non utilizzabili a causa della mancanza di qualità delle variabili e la risoluzione di dati inconsistenti.

Se è richiesto di accedere a informazioni disponibili su ASIA, si utilizza la precedente versione del registro (anno t - 1).

Come già illustrato, relativamente alla fase di eliminazione di unità duplicate, in *GSBPM* non esiste un sottoprocesso esplicito ma, per l'affinità del tipo di procedure, questa attività rientra in un sottoprocesso di integrazione finalizzata all'identificazione dei *record* duplicati.

5.1 *Integrated data* (finalizzata all'operazione di deduplica).

Indicatori calcolati: Percentuali di unità presenti più volte con gli stessi identificativi all'interno della stessa fonte.

A un'analisi approfondita, da quando l'acquisizione dall'Agenzia delle entrate è totalmente *on-line*, le duplicazioni corrispondono a *record* che vanno a sostituire quelli già presentati, a causa di successive correzioni.

13 RACLI è il Registro con informazioni su occupazione, retribuzioni, costo del lavoro e ore per la singola posizione lavorativa dipendente e relativa impresa (struttura di tipo *LEED - Linked Employer-Employee Database* con informazioni su individuo e impresa), del settore privato extra-agricolo.

Il successivo *step* consiste nel controllo della completezza e qualità delle variabili. L'obiettivo è ancora eliminare le unità i cui valori delle variabili rendono il *record* inutilizzabile. Si può parlare di una revisione e validazione dei dati delle fonti.

5.3 ***Review and validate***.

Indicatori calcolati: Percentuali di Unità eliminate perché non utilizzabili.

Infine, vengono applicate a ciascuna fonte delle procedure di controllo e correzione personalizzate, con l'obiettivo di risolvere eventuali inconsistenze all'interno della fonte.

5.4 ***Edit and impute***.

Indicatori calcolati relativi alle regole:

- Percentuale di regole di controllo violate almeno una volta;
- Percentuale unità con almeno una regola violata.

Indicatori calcolati relativi ai controlli:

- Tasso di imputazione (da 0 o *missing* a valore);
- Tasso di modificazione (da valore ad altro valore);
- Tasso di imputazione netta (da *missing* a valore);
- Tasso di cancellazione (da valore a *missing*);
- Tasso di non imputazione (valori o *missing* o 0 rimasti uguali);
- Tasso di valori *missing* non modificati (*missing* rimasti *missing*);
- Tasso di valori non *missing* non modificati (valori o 0 rimasti uguali).

Indicatori calcolati relativi alle unità controllate:

- Percentuale di modificazione: numero di unità da valore (non 0 o *missing*) ad altro valore (non 0 o *missing*) almeno una volta sul totale di unità imputate;
- Percentuale di imputazione (netta): numero di unità da *missing* a valore (non 0) almeno una volta sul totale di unità imputate;
- Percentuale di cancellazioni: numero di unità da valore (non 0) a *missing* almeno una volta sul totale di unità imputate;
- Percentuale di unità con valori *missing* non modificati sul totale unità controllate;
- Percentuale di unità con valori non *missing* o 0 non modificati sul totale unità controllate.

5.1 ***Integrated data*** (finalizzata al confronto con la popolazione *target*).

Successivamente, le unità di ciascuna singola fonte vengono *linkate* ad ASIA (marzo, anno $t + 2$). Si ha pertanto un vero e proprio passo di integrazione.

Indicatori calcolati: Tassi di Copertura (PRE-TRATTAMENTO) di ciascuna fonte rispetto allo stesso registro statistico ASIA.

Fase 3. Armonizzazione delle fonti.

Questa fase riguarda le analisi di confronto di variabili analoghe presenti in più fonti e la standardizzazione e ricodifica delle variabili. Così come tale, questa attività non sembra essere adeguatamente descritta in alcun sottoprocesso di *GSBPM*. In questa fase, si individuano e si trasformano gli *item* amministrativi per farli coincidere o al più avvicinare al significato statistico delle variabili economiche del regolamento SBS. In alcuni casi le analisi quindi possono essere interpretate come distanza tra i concetti amministrativi e quelli statistici.

Indicatori calcolati: Differenze percentuali tra variabili, Distanze tra distribuzioni.

Le variabili analoghe presenti in più fonti vengono rinominate per distinguerle nella fase di prioritizzazione nella scelta delle fonti. Vi è poi una fase di ricodifica. Il sottoprocesso di *GSBPM* è:

5.2 ***Classify and code***.

Indicatori calcolati: Percentuali di variabili desumibili senza riclassificazione.

Fase 4. Calcolo indicatori finali di copertura delle fonti.

Indicatori calcolati: Tassi di Copertura (POST-TRATTAMENTO) di ciascuna fonte rispetto allo stesso registro statistico ASIA.

Fase 5. Integrazione delle fonti.

La procedura di integrazione è deterministica e segue una scelta gerarchica delle fonti in base alla loro affidabilità, e tenendo conto di casi specifici per i quali vengono identificati degli errori sui dati che fanno derogare alla gerarchia sulle fonti.

5.1 ***Integrated data***.

Indicatori calcolati sulle unità:

- Percentuale di unità presenti in una fonte;
- Percentuale di unità presenti in più fonti;
- Percentuale di unità presenti in almeno una fonte;
- Percentuale di unità non presenti in alcuna fonte.

Indicatori calcolati sulle fonti:

- Percentuale di unità che derivano nel *dataset* integrato da una o da un'altra fonte.

6. Il *quality-framework* di *Statistics* Austria: studio di applicabilità al Sistema Integrato dei Registri Istat

Tra i vari approcci proposti in letteratura per la misurazione della qualità dei processi basati sull'utilizzo dei dati amministrativi, risulta interessante quello proposto dai ricercatori di *Statistics* Austria in occasione del Censimento 2011 (Asamer *et al.*, 2016; Lenk, 2008). Nella nostra analisi, se ne propone un'applicazione a un registro del SIR, per testare la possibilità di integrare il *framework* proposto nel paragrafo 5, con alcuni elementi aggiuntivi orientati anche alla definizione di misure di qualità riassuntive o sintetiche. Questo permetterebbe di definire ulteriori dimensioni di analisi dei processi utili a fornire informazioni aggiuntive per la definizione di misure di qualità degli *output*, oggetto di futuri sviluppi dell'approccio perseguito all'Istat.

6.1 Il *quality-framework* austriaco: i livelli e le iperdimensioni

Negli ultimi anni *Statistics* Austria, avendo condotto nel 2011 un censimento basato interamente su dati amministrativi, ha avvertito l'esigenza di dare avvio a un'analisi dei risultati attraverso la formulazione di un apposito *framework* che potesse permettere la valutazione in termini di qualità sia degli *output* delle variabili rilasciate, sia dei singoli *input* entrati ed elaborati nel processo dell'*output* stesso.

Nel dettaglio, per condurre il proprio censimento, *Statistics* Austria ha impiegato un sistema basato su più livelli. Al livello dei dati grezzi (*raw data*) il sistema colloca i cosiddetti *administrative registers* (che corrispondono alla definizione data nel sottoparagrafo 2.1), ossia fonti di natura amministrativa alimentate da numerosi enti esterni. Questi registri amministrativi sono stati utilizzati sia per fornire i valori delle variabili di interesse sia per la loro verifica: a tale proposito i *raw data* vengono infatti suddivisi rispettivamente in *base* e *comparison registers*. In questo caso, la classificazione delle tipologie dei registri è in funzione del loro ruolo nei processi di validazione delle variabili di interesse: i *base registers* forniscono il valore del dato di interesse in quanto reputato più corretto, mentre i *comparison registers* sono utilizzati per il confronto e la validazione di tali valori, secondo quello che viene definito "principio di ridondanza" (Lenk, 2008). Questa flessibilità nell'uso delle fonti è dovuta all'esperienza che l'istituto austriaco ha potuto sviluppare nel trattamento e nell'utilizzo delle stesse, per le quali è stata in grado di disporre di una gerarchia in base alla qualità e all'affidabilità dei dati acquisiti (Schnetzer *et al.*, 2015). Questa esperienza ha permesso inoltre all'istituto austriaco di definire indicatori di *input* dettagliati che presuppongono un forte rapporto di collaborazione con i fornitori.

Dopo l'acquisizione dei dati amministrativi il percorso per ottenere il dato finale si compone di diversi passaggi. Il primo è costituito dall'integrazione dei *dataset* di *input* in ipercubi tematici appartenenti a un *database* centrale (CDB, *Central DataBase*); successivamente l'imputazione dei dati mancanti permette di ottenere il *final data pool* (FDB), costituito dai registri statistici contenenti i dati utilizzati per il censimento. Il percorso di vita del dato, così definito, svolge un ruolo centrale nel *framework* di qualità proposto da *Statistics* Austria: infatti gli indicatori elaborati seguono l'intero percorso di costruzione del dato, dalla sua acquisizione fino al suo utilizzo per l'*output* finale, secondo una sequenza di passi

successivi che possono essere delineati attraverso tre livelli e quattro iperdimensioni di riferimento:

i. Livello dei dati grezzi (*raw data*);
ii. Livello del *dataset* di dati grezzi combinati (CDB, *Central DataBase*);
iii. Livello del *dataset* finale con imputazioni (FDP, *final data pool*).

a. Iperdimensione della documentazione: $\mathbf{HD^D}$;
b. Iperdimensione del pre-trattamento: $\mathbf{HD^P}$;
c. Iperdimensione del confronto con fonti esterne: $\mathbf{HD^E}$;
d. Iperdimensione dell'imputazione: $\mathbf{HD^I}$.

L'approccio austriaco, tenendo conto della specificità del contesto di analisi, propone una serie di indicatori per ciascun livello e dimensione sopra elencati con dettaglio sulla singola variabile. È importante sottolineare che la sequenza di passaggi per la costruzione degli indicatori segue le tre principali macro-fasi di un processo statistico: *input*, *process* e *output* (Brancato, 2018). In tal modo, con una certa approssimazione, la fase di *input* può essere posta in coincidenza con il livello dei dati grezzi, i quali non si riferiscono soltanto ai dati amministrativi acquisiti dai registri ma all'intero flusso di *input*, ovvero dati provenienti da indagini statistiche o persino altri registri. In questa prima fase di *input* l'indicatore finale è dato da una somma pesata degli indicatori elementari delle singole dimensioni, con un livello di dettaglio per registro e variabile.

Passando alla fase di *process*, i dati grezzi sono combinati nel *database* centrale e gli indicatori intermedi sono a livello di variabile e unità. In particolare per le variabili presenti in più registri si ottiene un indicatore intermedio dato da una combinazione degli indicatori riferiti ai singoli registri secondo la teoria probabilistica di Dempster-Shafer (Shafer, G., 1992), che rappresenta una generalizzazione della teoria bayesiana della probabilità. L'approccio probabilistico promosso da Dempster e Shafer si fonda sull'aggregazione di probabilità soggettive, interpretate come gradi di fiducia per l'accadimento di un determinato evento, in una probabilità complessiva. Nel *framework* austriaco tali probabilità sono identificate con i valori degli indicatori di qualità delle singole fonti. Questi costituiscono, secondo la suddetta interpretazione, il grado di fiducia che si può esprimere rispetto alla modalità assunta da un'unità della popolazione all'interno di una fonte: un valore elevato, proprio di fonti per cui la qualità è stata giudicata buona, determina un forte grado di fiducia che la modalità presente in quella fonte sia corretta. A questo punto è anche possibile calcolare un indicatore di confronto con una fonte esterna, se non lo si è fatto nella fase dei dati grezzi, e calcolare poi un indicatore sintetico complessivo.

Nell'ultimo passaggio, relativo alla fase di *output*, l'indicatore va a concentrarsi sulla qualità della procedura di imputazione. L'indicatore infatti combina la valutazione ottenuta nella fase precedente con un indicatore relativo all'imputazione. Alla fine di questa fase si otterrà per ciascuna variabile un valore univoco rappresentativo della sua qualità all'interno del sistema dei registri.

Scopo di questo lavoro è di valutare la generalizzazione dell'approccio austriaco applicandolo al Sistema Integrato dei Registri dell'Istat, tenendo conto delle differenze tra i relativi contesti.

6.2 Un'applicazione al SIR: l'esempio del Registro Base degli individui, delle famiglie e delle convivenze (RBI)

Il *quality-framework* austriaco è stato sviluppato da Asamer *et al.* (2016) sfruttando le informazioni disponibili nel sistema austriaco dei registri amministrativi, presi in considerazione specificatamente per il Censimento 2011. Nel presente studio si ipotizza un'applicazione di tale *framework* al RBI, uno dei registri statistici di base del SIR italiano (cfr. sottoparagrafo 2.2).

Il RBI è costituito da un insieme di individui estratti da SIM secondo determinati criteri di eleggibilità prefissati e sottoposti a pseudonimizzazione (sottoparagrafo 4.1). Per questi individui sono noti (a un prefissato livello di attendibilità) i valori relativi a un insieme di variabili sufficientemente invarianti nel tempo, le cosiddette variabili di eleggibilità (genere, data di nascita, luogo di nascita, data di decesso se valorizzata) e i valori di cittadinanza e grado di istruzione (variabili *core* mutabili nel tempo) (metodi riportati nel volume a cura di Di Zio e Vivio, 2019). Il RBI contiene quindi sia individui residenti ossia iscritti in anagrafe sia individui non residenti. Esistono diversi flussi che alimentano il RBI e che coinvolgono una o più fonti di dati determinando l'esigenza di gestire più processi contemporaneamente, uno per variabile o gruppo di variabili.

Lo studio è stato condotto ipotizzando di considerare due variabili: il sesso, una variabile *core* anagrafica immutabile nel tempo, e la cittadinanza, variabile *core* mutabile nel tempo, entrambe considerate secondo la definizione di Berka *et al.* (2010, 2012) come *multiple attribute*, perché più registri amministrativi forniscono informazioni sulla stessa variabile: il sesso è presente in 30 fonti mentre la cittadinanza è ottenuta a partire dalle informazioni riportate su 13 archivi.

In questa prima fase del lavoro, l'applicazione si è concentrata sul primo livello, quello relativo ai dati grezzi, che nel caso del RBI e delle variabili prese in considerazione, interessa i singoli archivi (e flussi nel caso del sesso) utilizzati nei processi di costruzione delle variabili.

Seguendo il modello austriaco sono state considerate tre iperdimensioni a livello di dati grezzi: quella relativa alla documentazione, l'iperdimensione del pre-trattamento e infine quella inerente al confronto delle fonti esterne.

Mentre *Statistics* Austria ha fatto affidamento anche sull'interazione con l'esterno per ricavare alcune informazioni di supporto al proprio *framework* (in particolare somministrando un questionario semistrutturato ai detentori dei registri amministrativi per ricavare i metadati riguardanti la fase di *input*), in questo studio si è optato per la valorizzazione degli strumenti e dei sistemi presenti in Istat: QRCA e SIDI-SIQual.

Si è navigato nei due sistemi, relativamente agli archivi e ai flussi coinvolti nei processi di costruzione delle variabili sesso e cittadinanza, nell'ottica di individuare tutta una serie di informazioni che potessero entrare nel calcolo degli indicatori di qualità estesi a tutte le variabili del Registro Base degli Individui.

Come descritto nel sottoparagrafo 3.1, la QRCA contiene la documentazione degli archivi amministrativi presenti in SIM. Il sistema SIDI-SIQual (cfr. sottoparagrafo 1.2) è stato utilizzato per reperire informazioni relativamente ai flussi demografici coinvolti nello studio della variabile sesso.

Per l'iperdimensione del pre-trattamento, anch'essa riguardante i *raw data*, è stato necessario attingere alla QRCA e a SIDI-SIQual per poter determinare, ad esempio, il numero dei *record* utilizzabili.

Infine è stata studiata l'iperdimensione del confronto con fonti esterne per la quale risulta necessario avviare una riflessione e un approfondimento.

Nei successivi sottoparagrafi vengono descritti, relativamente alle tre iperdimensioni prese in esame, gli indicatori e le misure proposte sulla base delle analisi svolte sulle informazioni disponibili o desumibili dai sistemi presenti in Istat, seguendo l'approccio proposto da Asamer *et al.* (2016).

6.2.1 Iperdimensione della documentazione

L'iperdimensione della documentazione (HD^D) racchiude gli aspetti legati alla disponibilità e all'accessibilità dei metadati, inclusa la presenza di documentazione adatta a valutare le fonti che confluiscono nei registri statistici. Gli indicatori relativi a questa iperdimensione riguardano infatti le fasi a monte del processo produttivo, in particolare l'acquisizione dei dati grezzi. È utile a questo punto evidenziare che nell'ambito del presente studio viene proposta un'associazione tra le iperdimensioni identificate da *Statistics* Austria e le dimensioni della qualità definite da Eurostat (Eurostat, 2003). Pertanto, in accordo a questo modello, per l'iperdimensione della documentazione è stato ritenuto opportuno associarvi le dimensioni della confrontabilità, della chiarezza e della tempestività e puntualità, ciascuna con i relativi indicatori che sono stati adattati o riformulati dalla proposta originaria del *framework* per i registri austriaci. Le dimensioni proposte in riferimento all'iperdimensione della documentazione e gli indicatori associati sono riportati nel Prospetto 6.1.

La dimensione della confrontabilità concerne la possibilità da parte degli utenti di poter effettuare confronti in serie storica sui dati di interesse; circostanze quali interruzioni nelle serie temporali o cambiamenti nelle definizioni adottate influiscono negativamente su questa dimensione. Per essa sono stati quindi introdotti due indicatori, uno per la valutazione delle interruzioni della serie e l'altro per la valutazione della disponibilità dello storico dei tracciati *record*; il primo di questi indicatori è calcolabile su entrambi i sistemi dell'Istat a cui si è fatto riferimento per la ricerca di metadati, ovvero QRCA e SIDI-SIQual; il secondo indicatore è invece calcolabile solo per QRCA, pur con l'accortezza che al momento nel sistema viene reso disponibile solo l'ultimo tracciato in ordine temporale. Entrambi gli indicatori sono dicotomici, prevedendo una risposta di tipo "sì/no" alla presenza delle

informazioni richieste; tuttavia non è da escludere la possibilità di ottenere ulteriori tracciati *record* con richieste apposite, per cui nel relativo indicatore questa modalità di risposta è stata affiancata alle altre.

La dimensione della chiarezza è stata esplicitata nell'indicatore sull'adeguatezza della documentazione, più complesso dei precedenti in quanto la valutazione della documentazione presenta necessariamente alcuni aspetti di soggettività. In questo caso si è scelto di non usare un indicatore dicotomico, bensì di valutare questo indicatore su una scala da 0 a 5 sia per le informazioni su QRCA sia su SIQual.

Infine all'iperdimensione della documentazione possono essere ricondotte le caratteristiche legate alla temporalità dei dati, rappresentate dalle dimensioni della tempestività e della puntualità. Per questi indicatori si è fatto riferimento alle definizioni prodotte da Eurostat, che individuano la tempestività come differenza in giorni tra la diffusione dei dati e il loro periodo di riferimento e la puntualità come differenza tra la data di diffusione effettiva e quella programmata.

Naturalmente, nel caso delle fonti utilizzate dai registri la diffusione non sarà quella destinata al pubblico ma ai titolari dei registri. Anche per questi indicatori si è scelto, per esigenze di standardizzazione e di continuità con l'indicatore precedente, di procedere a una valutazione basata su una variabile ordinale da 0 a 5 piuttosto che su un calcolo esatto del numero di giorni intercorsi tra le specifiche date.

Una volta calcolati tutti gli indicatori illustrati, per ottenere il punteggio complessivo per questa iperdimensione, gli indicatori proposti (in questo caso 5) vengono combinati e il risultato restituirà un indicatore finale standardizzato che varia tra 0 e 1.

Prospetto 6.1 – Indicatori e rispettive misure individuate per l'iperdimensione della documentazione (a)

Dimensione	Indicatore	QRCA sì/no	Note QRCA	SIDI-SIQual sì/no	Note SIDI-SIQual	Misurazione	Valore
Confrontabilità	Interruzioni nella serie temporale	sì		sì		0 - no 1 - sì	Punteggio
	Disponibilità dello storico dei tracciati *record*	sì	Solo ultimo disponibile		Approfondire la disponibilità	0 - no 1 - richiesta *ad hoc* 2 - sì	Punteggio/2
Chiarezza	Adeguatezza della documentazione	sì	Dettagliabile per livelli diversi di esigenza	sì	Dettagliabile per livelli diversi di esigenza	Scala ordinale 0 - pessima 5 - ottima	Punteggio/5
Tempestività e Puntualità	Puntualità	sì	Stato della puntualità, giorni di ritardo o anticipo	sì		Scala ordinale 0 - molto in ritardo 5 - puntuale/anticipo	Punteggio/5
	Tempestività	sì	Calcolabile per l'ente e per Istat	sì		Scala ordinale 0 - poco tempestivo 5 - molto tempestivo	Punteggio/5

Fonte*:* Schema prodotto dagli autori
(a) *Hyperdimension Documentation* HD^D Somma indicatori/5.

6.2.2 Iperdimensione del pre-trattamento

L'iperdimensione del pre-trattamento (HD^P) riguarda gli errori formali presenti nei dati grezzi e, nello specifico, permette di valutare la dimensione di qualità dell'accuratezza (Eurostat, 2003). È importante sottolineare che non tutti gli indicatori sono direttamente calcolabili con le informazioni già disponibili nei sistemi QRCA e SIQual, anche se sussiste la possibilità di attivare richieste *ad hoc* nel caso in cui manchino informazioni.

Per le fonti le cui informazioni sono contenute nella QRCA, è possibile controllare gli errori dovuti a:

- mancanza di codici identificativi;
- *record* non valorizzati;
- valori errati;
- valori fuori *range*.

Invece, per le fonti le cui informazioni sono contenute in SIDI-SIQual, è possibile misurare questo tipo di errori:

- unità non eleggibili;
- unità non rispondenti;
- valori imputati.

Il risultato finale della fase di pre-elaborazione (HD^P) è dato dal rapporto tra i *record* utilizzabili e il numero totale di *record*. Questa procedura viene eseguita per singolo registro o, se disponibile, per singola variabile.

In questa fase di adattamento del *framework* austriaco al RBI si è ritenuto opportuno tenere distinto dall'indicatore HD^P(SIQual) l'indicatore denominato tasso di imputazione che è presente nel sistema SIDI-SIQual solo per alcune variabili, e calcolare due indicatori HD^P_1(SIQual) e HD^P_2(SIQual) (Prospetto 6.2). Questo tipo di indicatore rientrerebbe nella fase finale del *framework* (livello del *dataset* finale con imputazioni), tuttavia, riferendosi in SIDI-SIQual ai dati grezzi, è possibile calcolarlo nella valutazione di qualità di questa fase. In una versione successiva di specializzazione del *framework* si può ipotizzare di far rientrare i valori imputati come elemento del numeratore dell'indicatore finale HD^P(SIQual).

Nel Prospetto 6.2 sono riportati gli indicatori individuati e le ipotesi di misurazione per ottenere, come per il modello austriaco, indicatori a livello di archivio e/o variabile.

Prospetto 6.2 – Indicatori e rispettive misure individuate per l'iperdimensione del pre-trattamento (a)

Dimensione	Indicatore	QRCA sì/no	Note QRCA	SIDI-SIQual sì/no	Note SIDI-SIQual	Misurazione	Valore
Accuratezza	Numero di *record* valorizzati	sì	Per variabile. Il numero di osservazioni è derivabile			Numero osservazioni (A)	
	Numero di *record* con codice identificativo mancante	sì	Disponibile su richiesta			Numero osservazioni senza ID (B)	
	Numero di *record* con valore fuori *range*	sì	Presente per variabili categoriche la cui classificazione amministrativa viene fornita dal titolare della fonte. Per le altre variabili sono necessarie elaborazioni *ad hoc*			Numero osservazioni fuori *range* (C)	
	Totale unità			sì		Totale unità (D)	
	Unità non eleggibili			sì	Calcolabile	Unità non eleggibili (E)	
	Unità non rispondenti			sì	Calcolabile	Unità non rispondenti (F)	
	Rapporto di non imputazione			sì	Quando presente per alcune variabili. Altrimenti su richiesta	Tasso di imputazione/100 (G)	

Fonte: Schema prodotto dagli autori
(a) *Hyperdimension Pre-processing* HD^P (QRCA) (A - B - C)/A,
Hyperdimension Pre-processing HD^P_1(SIDI-SIQual) (D - E - F)/D,
Hyperdimension Pre-processing HD^P_2(SIDI-SIQual) G.

6.2.3 Iperdimensione del confronto con fonti esterne

La terza iperdimensione, quella relativa al confronto con le fonti esterne (HD^E), presuppone l'esistenza di una o più fonti esterne al RBI con cui operare il confronto calcolando i valori che risultano coerenti. Questa iperdimensione, a differenza delle altre, può riguardare tutti e tre i livelli: *raw data*, dati grezzi combinati e *dataset* finale con imputazioni.

Nello specifico, a livello di dati grezzi, si è ipotizzato di seguire il lavoro di Asamer *et al.* (2016) e valutare la dimensione della Coerenza mediante il calcolo del rapporto tra il numero dei valori coerenti e il totale dei valori controllati, lasciando aperta la questione sull'individuazione della fonte esterna da utilizzare per effettuare il confronto. Come è indicato nel Prospetto 6.3, la decisione su quale fonte usare è condizionata dalle caratteristiche strutturali della variabile considerata, dal tipo di procedura usata per stimare la variabile e dalle eventuali indicazioni di esperti. La scelta potrebbe ricadere sul *Master Sample* (MS[14]),

14 *Master Sampl*e: è la base dati ottenuta mediante due rilevazioni campionarie del Censimento della Popolazione, una componente da lista e una componente areale, che contribuiscono alla determinazione della popolazione fino a specifici livelli di disaggregazione territoriale e tematica.

sfruttando la capacità informativa del dato rilevato. L'approccio modulare del *framework* austriaco permetterebbe di effettuare il confronto con il MS a valle della costruzione del RBI. Da quanto visto in questa fase preliminare di studio è emersa l'esigenza di avviare un approfondimento e una riflessione accurata su questa iperdimensione.

Prospetto 6.3 – Indicatore e rispettiva misura individuata per l'iperdimensione del confronto con fonti esterne

Dimensione	Indicatore	Misurazione	Note
Coerenza	Coerenza con le fonti esterne	Numero di valori coerenti/totale dei valori controllati	La scelta della fonte esterna è condizionata dalle caratteristiche strutturali della variabile considerata, dal tipo di procedura usata per stimare la variabile e dalle eventuali indicazioni di esperti

Fonte: Schema prodotto dagli autori

6.3 Valutazione dell'applicabilità e sviluppi futuri

In questo documento è stato presentato un primo tentativo di adattare il *framework* di Asamer *et al.* (2016) alle peculiarità del processo di produzione del RBI, al fine di identificare le effettive fonti di errore e le corrispondenti misure di qualità sia sequenzialmente nelle fasi di *input*, *process* e *output* (Brancato, 2018), sia attraverso gli ulteriori moduli delle iperdimensioni.

In un'ottica di utilizzatore del modello, si è adottato un approccio di analisi di tipo *bottom-up*, ossia, come descritto in precedenza, si è valutata l'immediata disponibilità di dati e indicatori elementari che potessero essere utili nella costruzione del *framework*. Bisogna sottolineare che questo primo studio si è concentrato sulle iperdimensioni della fase di *input*, poiché gli strumenti per la costruzione degli indicatori, QRCA e SIDI-SIQual, fanno riferimento alla qualità delle fonti di dati grezzi.

Come già descritto nei sottoparagrafi precedenti, non sempre le informazioni sulla qualità sono immediatamente disponibili così come definite nel lavoro degli austriaci, questo perché la QRCA e il SIDI-SIQual, progettati in Istat secondo obiettivi e tempistiche diverse, non sono stati pensati per coprire tutti gli aspetti di valutazione di un registro multifonte come RBI. Infatti, il lavoro dell'istituto austriaco desume molti indicatori da un questionario di valutazione compilato *ad hoc*. Quindi, per avere un risultato simile a quello austriaco, sarebbe necessario definire ulteriori modalità di reperimento delle informazioni utili allo scopo.

Un primo adattamento effettuato è stato quello di dividere le iperdimensioni in sottodimensioni che potessero identificare meglio la natura degli indicatori elementari: confrontabilità, chiarezza, tempestività e puntualità, accuratezza e coerenza. Questo permette di individuare, dentro ogni iperdimensione, sia l'aspetto che dovrebbe essere migliorato sia la natura dell'errore più presente.

L'applicazione del *framework* austriaco, che è stato preso come punto di riferimento, è possibile solo in parte; si tratta di una procedura generalizzata che si basa su un approccio sequenziale (*input*, *process* e *output*) di diversi moduli (iperdimensioni) e ciò lo rende utilizzabile per vari scopi. Sicuramente la struttura di valutazione proposta dagli austriaci, che segue il flusso dei dati lungo tutta la filiera produttiva del sistema dei registri, ben si

adatta a quella del RBI ma, sia per il modo con cui è stato costruito sia per gli scopi specifici del RBI, servono necessari adattamenti negli indicatori elementari.

I passi successivi dello studio saranno quelli di estendere l'applicazione del *framework* ai rimanenti passi di *process* e *output* così da avere una panoramica completa degli aspetti da migliorare nella personalizzazione del *framework*.

Va notato che l'identificazione e la successiva eliminazione delle cause di errore rappresenta la base per il miglioramento sistematico e continuo del processo di produzione del RBI. Inoltre, grazie alla disponibilità di un sistema di valutazione sarebbe possibile analizzare sia la qualità dei dati sia il processo di produzione del registro in una prospettiva longitudinale. In seguito, sulla base dell'adeguamento del *framework* preso in considerazione, potrebbe essere sviluppato un rapporto, a fini di documentazione e diffusione, sulla qualità del registro, nonché estendere lo studio anche agli altri registri del SIR, in modo da affrontare ulteriori tematiche e criticità non riscontrate nell'analisi del solo RBI.

L'evoluzione dello studio potrebbe infine essere quella di integrare l'approccio austriaco con altre esperienze presenti in letteratura sia in termini strutturali (Zhang, 2012; Reid *et al.*, 2017), alla base anche di esperienze nate e sviluppatesi internamente all'Istat (Luzi *et al.,* 2017), sia per la scelta di indicatori elementari più adeguati al caso studio (Daas and Ossen, 2011; Reinert and Stoltze, 2016).

7. Conclusioni e passi futuri

Questo lavoro ha descritto i risultati ottenuti nei 6 mesi di lavoro del GdL finalizzato alla definizione di un sistema di documentazione dei processi e di monitoraggio della qualità del SIR. Quanto portato a termine costituisce l'infrastruttura fondamentale per sviluppare tale sistema, tuttavia vi sono ancora attività necessarie e propedeutiche per la sua implementazione finale.

In particolare, uno sforzo è stato effettuato per sistematizzare i concetti riguardanti le tipologie di dati e di registri sottostanti il SIR e verificarne la coerenza. Durante i lavori del gruppo, si è evidenziato come questa sistematizzazione non sia ancora patrimonio informativo consolidato dell'Istat, rimane quindi la necessità di divulgare ulteriormente la terminologia proposta. Possibili attività future in tale senso potranno essere l'aggiornamento del glossario e il relativo *iter* per la sua ufficializzazione, nell'ottica di pervenire a una mappatura completa degli oggetti che compongono il SIR.

Relativamente alla documentazione a supporto del monitoraggio e della valutazione della qualità, sono stati analizzati vari approcci in letteratura, dai modelli standard per la documentazione dei metadati (*GSIM*, *GSBPM*), ai *framework* generali e teorici sulla produzione statistica mediante dati secondari come il ciclo produttivo a due fasi proposto da Zhang (Zhang, 2012). Inoltre sono stati analizzati gli indicatori attingendo da varie fonti: indicatori standard proposti nell'ambito del Sistema Statistico Europeo (Eurostat, 2020) e in vari ambiti collaborativi a livello europeo, come il progetto *Komuso – Quality of multisource statistics* (Brancato, 2018); approcci metodologici come il modello teorico sugli errori proposto da Zhang; esperienze empiriche condotte all'Istat (Luzi *et al.*, 2017) e da altri istituti nazionali di statistica, come l'approccio sul calcolo e la riduzione degli indicatori in un processo specifico proposto da Asamer *et al.* (2016). In base a questa ricognizione è stato identificato un *framework* armonizzato a livello internazionale e personalizzato al contesto produttivo dell'Istat. Questo include tutti i sottoprocessi rilevanti del *GSBPM* nella gestione dell'aggiornamento a regime dei registri del SIR. Inoltre, è stato identificato un primo set di indicatori e misure di qualità utili per il monitoraggio e la valutazione della qualità, con un *focus* particolare sugli indicatori che si possono calcolare per i principali sottoprocessi e che contribuiscono alla valutazione della dimensione dell'accuratezza. Sono stati identificati e proposti gli elementi di documentazione e le misure di qualità del processo per i singoli registri del SIR.

Quanto ottenuto rappresenta la base per futuri sviluppi. In primo luogo gli elementi iniziali ottenuti dovranno essere sperimentati a livello quantitativo su un sottoinsieme di registri del SIR, al fine della loro validazione definitiva. Inoltre, il lavoro dovrà essere integrato con ulteriori indicatori di qualità del processo, indicatori di coerenza tra variabili che sono in relazione tra loro appartenenti a diversi registri del SIR e misure relative a tutte le dimensioni della qualità ampliando l'attenzione alla qualità dell'*output*. Al momento gli indicatori sono definiti come singole misure che riflettono la qualità di sottoprocessi o caratteristiche dell'*output*, è stata sviluppata una prima riflessione sulla costruzione di misure sintetiche personalizzando l'approccio di Asamer *et al*. (2016) alla situazione Istat, ma tale riflessione non è ancora sufficientemente matura. Sarà quindi importante proseguire

il lavoro in tale senso per arrivare a definire l'utilità e l'opportunità di avere tali misure sintetiche.

Infine, il sistema di documentazione della qualità e degli indicatori per i registri del SIR dovrà essere sviluppato in modo integrato ed efficiente rispetto ai relativi sistemi di documentazione già esistenti all'Istat o alle loro evoluzioni. In tal senso, sono già state identificate le principali interconnessioni che il sistema di documentazione del SIR dovrà avere con gli altri sistemi interni all'Istat (Psn, ARCAM, QRCA, SIDI-SIQual, SUM), ma ne rimangono da approfondire le necessarie relazioni che il nuovo sistema dovrà avere con quelli esistenti. Su questo punto potrà essere di supporto anche una futura attività di modellazione dell'informazione mediante l'approccio delle ontologie.

Gli sviluppi futuri delineati saranno propedeutici alla realizzazione del sistema per la memorizzazione, l'accesso e la gestione dei metadati e degli indicatori di qualità del SIR, attraverso la sua progettazione tecnica e il suo sviluppo informatico.

Appendice - Formule di calcolo e interpretazione degli indicatori introdotti nelle schede del paragrafo 5.3

Di seguito, per ogni indicatore di ogni sottoprocesso sono riportati il nome, la formula di calcolo e l'interpretazione. Nel caso in cui la formula sia banale e desumibile dal nome dell'indicatore stesso, questa viene omessa.

Prospetto A1 – Sottoprocesso 1: Acquisizione dei dati della fonte

Indicatore	Formula	Interpretazione
1.1.Numero di *record* acquisiti nella fornitura rispetto allo stesso numero nella fornitura precedente.		Un numero di *record* molto maggiore o inferiore alla fornitura precedente è un campanello di allarme su possibili errori o omissioni nei dati trasmessi.
1.2. Documentazione delle variazioni nella struttura dei dati e nei metadati.	Documentazione delle variazioni nei tracciati, nelle definizioni delle unità statistiche e delle variabili, nelle classificazioni utilizzate, nei nomi e nei formati delle variabili.	Utile per comprendere la stabilità e le variazioni dei dati.

Fonte*:* Schema prodotto dagli autori

Prospetto A2 – Sottoprocesso 2: Integrazione

Indicatore	Formula	Interpretazione
2.1. Percentuale di unità duplicate.	Numero di *record* duplicati rispetto a tutte le variabili sul totale dei *record*, per cento.	Costituisce una misura di qualità del *dataset* fornito, anche se le duplicazioni possono essere facilmente rimosse.
2.2. Tasso di *link* (o *match rate*).	Numero di *record* abbinati tra due *dataset* sul totale dei *record* presenti nel *dataset* più numeroso, per cento.	Si interpreta alla luce dei metadati sulle unità contenute nei due *dataset* e della copertura dei singoli *dataset*: se i *dataset* appartengono alla stessa popolazione e le popolazioni sono adeguatamente rappresentate, allora il tasso di *link* dovrebbe essere molto vicino a 100.
2.3. Tasso di falsi *link.*	Numero di *record* abbinati ma corrispondenti a unità diverse su totale dei *record* abbinati, per cento.	Può essere stimato solo attraverso indagini *ad hoc*, esperimenti, verifica manuale sugli abbinamenti, modelli di stima che utilizzano un campione di dati noti. L'abbinamento di *record* che appartengono a unità diverse è un errore potenzialmente molto grave in quanto si propaga in errori di misurazione su tutte le variabili assegnate all'unità incorrettamente considerata come la medesima.
2.4. Tasso di falsi *non-link.*	Numero di *record* non abbinati ma corrispondenti alle stesse unità su totale dei *record* corrispondenti a veri abbinamenti, per cento.	Può essere stimato solo attraverso indagini *ad hoc*, esperimenti, verifica manuale sugli abbinamenti, modelli di stima che utilizzano un campione di dati noti. Non abbinare, e quindi considerare come unità differenti, *record* che in realtà appartengono alla stessa unità può causare un errore di copertura e quindi un'alterazione della numerosità della popolazione in studio.
2.5. Tasso di dati mancanti o errati nella variabile di *linkage.*	Numero di *record* con valore mancante o inammissibili per la variabile di *linkage* su totale *record.*	La qualità della variabile di *linkage* è un indicatore di base per valutare la bontà dell'abbinamento. Ovviamente chiavi di *linkage* non disponibili o errate causano la non *linkabilità* o l'abbinamento errato di dati in *dataset* differenti.
2.6. Distanze tra distribuzioni sulle variabili rilevanti nel *dataset* integrato e nei *dataset* di *input.*	Indici di distanza tra distribuzioni di variabili categoriche e continue (Hellinger, Kolmogorov-Smirnov) tra dati relativi al *dataset* integrato e ai *dataset* di *input*.	In base alle ipotesi di base sulle distribuzioni di determinate variabili, questo indicatore può consentire di comprendere se l'abbinamento ha portato a conseguenze non attese. Può essere utile calcolarlo su specifiche sottopopolazioni.

Fonte*:* Schema prodotto dagli autori

Prospetto A3 – Sottoprocesso 3: Codifica

Indicatore	Formula	Interpretazione
3.1. Tasso di efficacia della codifica manuale.	Numero di valori codificati manualmente sul totale dei valori da codificare manualmente, per cento.	Più l'indicatore è vicino a cento, maggiore è l'efficacia della codifica manuale.
3.2. Tasso di efficacia della codifica interattiva.	Numero di valori codificati interattivamente sul totale dei valori da codificare interattivamente, per cento.	Più l'indicatore è vicino a cento, maggiore è l'efficacia della codifica interattiva.
3.3. Tasso di efficacia della codifica automatica.	Numero di valori codificati automaticamente sul totale dei valori da codificare automaticamente per cento.	Più l'indicatore è vicino a cento, maggiore è l'efficacia della codifica automatica.
3.4. Indicatori di carico sui codificatori: numero di codificatori.	Numero di codificatori.	Il numero di codificatori deve essere sufficiente in modo da non imporre un elevato carico di attività su ciascuno e nello stesso tempo non troppo elevato da rendere difficile la formazione e il monitoraggio.
3.5. Indicatori di carico sui codificatori: *workload*.	Numero medio di *item* da codificare per codificatore.	Il numero di *item* da codificare deve essere sufficientemente ampio da permettere al codificatore di diventare esperto e nello stesso tempo non troppo elevato per non imporre un elevato carico di lavoro e un effetto di stanchezza.
3.6. Tasso di mancata codifica.	Numero di valori non codificati su numero di valori sottoposti a codifica per cento.	Valori prossimi a 0 riflettono una buona *performance* della procedura di codifica in generale, anche se l'indicatore non permette di comprendere se i valori siano stati codificati in modo corretto.

Fonte*:* Schema prodotto dagli autori

Prospetto A4 – Sottoprocesso 4: Validazione dei microdati

Indicatore	Formula	Interpretazione
4.1. Percentuale di unità presenti più volte con gli stessi identificativi all'interno della stessa fonte.	Numero unità con identificativo ripetuto rispetto al totale delle unità.	Se nella fonte non è previsto che l'unità sia presente più volte, rappresenta possibili errori negli identificativi e quindi una misura della qualità del *dataset* forniti.
4.2. Percentuale di unità eliminate perché non utilizzabili.	Numero di unità cancellate dal *dataset* rispetto al totale delle unità.	Riflette l'entità dell'informazione che viene eliminata perché di scarsa qualità e quindi di scarsa utilità.
4.3. Numero di unità con almeno un valore mancante.	Numero di unità.	Quando il *record* corrisponde all'unità, identifica l'ammontare dei *record* con un valore mancante.
4.4. Tasso di mancata risposta parziale	Numero di valori osservati su valori dovuti per cento.	Rappresenta l'ammontare di informazione mancante a livello di singole variabili.
4.5. Numero di variabili che falliscono almeno una regola di edit per tipo di regola (*soft/hard*).	Numero di variabili che violano almeno una regola.	L'errore viene detto *hard* quando la violazione della regola individua con certezza un errore, viene detto soft quando i controlli sono costruiti su assunzioni basate sulle conoscenze sulle variabili in oggetto.
4.6. Percentuale unità con almeno una regola violata (*edit failure rate*).	Numero di unità che violano almeno una regola rispetto al totale.	È una misura della non qualità del *dataset*, in quanto maggiore è la quota di *record* che violano almeno una regole, minore è la qualità dei dati.

Fonte*:* Schema prodotto dagli autori

Prospetto A5 – Sottoprocesso 5: Controllo e correzione

Indicatore	Formula	Interpretazione
Misure di qualità relative ai dati grezzi:		
5.1. Numero di unità con almeno un valore mancante.	Numero di unità.	Uguale all'indicatore 4.3.
5.2. Numero di unità che falliscono almeno una regola di *edit* (*edit failure rate*).	Numero di unità che violano almeno una regola rispetto al totale.	Uguale all'indicatore 4.6.
5.3. Tasso di mancata risposta parziale.	Numero di valori osservati su valori dovuti per cento.	Uguale all'indicatore 4.4.
Misure di qualità relative alla procedura di imputazione:		
5.4. Tasso di imputazione.	Numero di valori imputati/ numero di *record*, per cento.	Riflette quanti valori sono stati imputati, ma non con quale entità.
5.5. Tasso di imputazione per variabile.	Numero di valori imputati su numero di *record*, per cento	Riflette quanti valori sono stati imputati, ma non con quale entità.
5.6. Tasso di modificazione per variabile.	Numero di valori modificati da codice a codice su numero di *record*, per cento.	Riflette quanti valori sono stati modificati, ma non con quale entità.
5.7. Tasso di imputazione netta per variabile.	Numero di valori modificati da *blank* a codice su numero di *record*, per cento.	Riflette quanti valori sono stati imputati (letteralmente assegnato un valore quando mancante), ma non con quale entità.
5.8. Tasso di cancellazione per variabile.	Numero di valori modificati da codice a *blank* su valori passibili di imputazione, per cento.	Riflette l'ammontare di dati di cattiva qualità, tanto da dover essere cancellati.
5.9. Tasso di imputazione ponderato .	Somma dei valori imputati sul totale stimato, per cento.	L'indicatore è applicabile solo a variabili quantitative. Riflette quanto pesano sul totale di una certa variabile le imputazioni effettuate.

Fonte*:* Schema prodotto dagli autori

Prospetto A6 – Sottoprocesso 6: Validazione dei microdati

Indicatore	Formula	Interpretazione
6.1. Numero di unità derivate o trasformate.	Numero di unità.	Rappresenta l'ammontare di unità ottenute come prodotto di una procedura di trasformazione o derivazione. Può essere utile confrontarlo con l'ammontare delle unità di partenza o di altre fonti per valutazioni di coerenza.
6.2. Tasso di copertura delle unità derivate o trasformate.	Numero di unità derivate o trasformate sul totale delle unità che fanno parte della popolazione obiettivo.	Riflette quanto il processo di derivazione o trasformazione è in grado di ricostruire tutti i segmenti della popolazione obiettivo, è necessario conoscere il totale della popolazione obiettivo.
6.3. Indicatori di confronto su variabili tra unità derivate e non derivate.	Misure di confronto tra medie, confronto tra distribuzioni, eccetera.	Nell'ipotesi che le due popolazioni relative alle unità derivate e non derivate provengano dalla stessa popolazione obiettivo, consente di valutare se il processo di derivazione ha portato a distorsioni nella distribuzione di alcune variabili.

Fonte*:* Schema prodotto dagli autori

Riferimenti bibliografici

Asamer, E.-M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk, M. Moser, and H. Rechta. 2016. "Quality assessment for register-based statistics - Results for the Austrian census 2011". *Austrian Journal of Statistics*, Volume 45, N. 2: 3-14.

Bellitti, G., e C. Colasanti (*a cura di*). 2021. "Manuale sui principali adempimenti in materia di trattamento di dati personali: il caso dell'Istat". *Letture statistiche - Metodi*. Roma, Italia: Istat. https://www.istat.it/it/archivio/259025.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer. 2012. "Combination of evidence from multiple administrative data sources: quality assessment of the Austrian register-based Census 201". *Statistica Neerlandica*, Volume 66, Issue 1: 18-33.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer. 2010. "A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011". *Austrian Journal of Statistics*, Volume 39, N. 4: 299-308.

Brancato, G. (*Ed.*). 2018. "Quality Guidelines for Multisource Statistics (QGMSS). Version 0.8.1". *ESSnet on Quality of Multisource Statistics - KOMUSO*. Specific Grant Agreement N. 2 (SGA - 2), Work Package 1: "Guidelines on the quality of multisource statistics", Deliverable 5. Luxembourg: Eurostat.
https://ec.europa.eu/eurostat/cros/system/files/wp1_guidelines_-_v0_8_1.pdf.

Brancato, G., A. Boggia, F. Barbalace, F. Cerroni, S. Cozzi, G. Di Bella, M. Di Zio, D. Filipponi, O. Luzi, P. Righi, e M. Scanu. 2016. "Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi. Versione 1.1. Agosto 2016". *Linee Guida per la Qualità*. Roma, Italia: Istat. https://www.istat.it/it/metodi-e-strumenti/strumenti-per-la-qualit%C3%A0/linee-guida.

Brancato, G., R. Carbini, C. Pellegrini, M. Signore, and G. Simeoni. 2006. "Assessing quality through the collection and analysis of standard quality indicators: the Istat experience". Paper presented at the *European Conference on Quality in Survey Statistics - Q2006*. Cardiff, UK, 24-26 April 2006.

Brancato, G., C. Pellegrini, G. Simeoni, and M. Signore. 2004. "Standardising, Evaluating and Documenting quality: the implementation of Istat information system for survey documentation – SIDI". Paper presented at the *European Conference on Quality and Methodology in Official Statistics - Q2004*. Mainz, Germany, 24-26 May 2004.

Cerroni, F., G. Di Bella, and L. Galiè. 2014. "Evaluating administrative data quality as input of the statistical production process". *Rivista di statistica ufficiale/Review of official statistics*, N. 1-2/2014: 117-146. Roma, Italy: Istat. https://www.istat.it/en/archivio/271641.

Costanzo, L., G. Di Bella, V. Talucci, M. Vignola, R. van de Laar, J.H. Pereira, S. Rodrigues, J. Darke, and A. Pritchard. 2013. "Admin Data Glossary. Definitions adopted for certain terms related to the use of administrative data for producing business statistics". *Deliverable 1.1 (31/1/2013). ESSnet Admin Data, Work Package 1: Overview of MSs' Existing Practices in the Uses of Administrative Data for Business Statistics*. Brussels, Belgium, and Luxembourg: European Commission.

Curatolo, S, V. De Giorgi, F. Oropallo, A. Puggioni, and G. Siesto. 2016. "Quality analysis and harmonization issues in the context of 'Frame SBS'". *Rivista di statistica ufficiale/Review of official statistics*, N. 1/2016: 15-46. Roma, Italy: Istat. https://www.istat.it/en/archivio/271235.

Daas, P., and S. Ossen (*Eds.*). 2011. "Report on methods preferred for the quality indicators of administrative data sources". *Deliverable 4.2 prepared for the BLUE-Enterprise and Trade Statistics (BLUE-ETS)*. Brussels, Belgium, and Luxembourg: European Commission.

Di Bella, G. (*a cura di*). 2021. "Il sistema di documentazione dei dati amministrativi in Istat". *Letture Statistiche – Metodi*. Roma, Italia: Istat. https://www.istat.it/it/archivio/263001.

Di Zio, M., e R. Vivio (*a cura di*). 2019. "La stima della cittadinanza nel Registro Base degli Individui tramite integrazione di fonti amministrative". *Letture Statistiche - Metodi*. Roma, Italia: Istat. https://www.istat.it/it/archivio/234307.

Eurostat. 2020. "European Statistical System handbook for quality and metadata reports. 2020 Edition". *Manuals and Guidelines*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-19-006.

Eurostat. 2003. "Item 4.2: Methodological Documents - Definition of Quality in Statistics". *Working Group "Assessment of Quality in Statistics"*. Sixth meeting, Luxembourg, 2-3 October 2003. https://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf.

Istituto Nazionale di Statistica - Istat. 2017. "Il sistema integrato dei registri. Architettura informativa e metodologica, Versione 1.0 del 20 dicembre 2017. *Documento tecnico interno*. Roma, Italia: Istat.

Istituto Nazionale di Statistica – Istat. 2016*a*. "Il Sistema Integrato dei Registri nell'ambito del processo di modernizzazione, Versione 0.1 dell'8 febbraio 2016". *Documento tecnico interno*. Roma, Italia: Istat.

Istituto Nazionale di Statistica – Istat. 2016*b*. "La progettazione del Sistema Integrato dei Registri, Versione 0.2 dell'1 agosto 2016". *Documento tecnico interno*. Roma, Italia: Istat.

Istituto Nazionale di Statistica – Istat. 2016*c*. *Il Programma di Modernizzazione dell'Istat*. Roma, Italia: Istat. https://www.istat.it/it/files/2010/12/Programma_modernizzazione_Istat2016.pdf.

Lenk, M. 2008. *Methods of Register-based Census in Austria*. Wien, Austria: Statistik Austria.

Luzi, O., F. Rocci, R. Seananzo, and R. Varriale. 2017. "A quality evaluation framework for the statistical register Frame-SBS". *Rivista di statistica ufficiale/Review of official statistics*, N. 1-2-3/2017: 67-85. Roma, Italy: Istat. https://www.istat.it/en/archivio/271229.

Luzi, O., F. Rocci, and R. Varriale. 2018. "Quality evaluation of statistical processes based on administrative data: a new version of the TSE approach". Paper presented at the *European Conference on Quality in Official Statistics - Q2018. Kraków*, Poland, 26-29 June 2018.

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.

Reinert, R., and P.T. Stoltze (*Eds.*). 2016. "Work Package 1. Checklist for Evaluating the Quality of Input Data". *ESSnet KOMUSO - Quality in Multisource Statistics*. Specific Grant Agreement N. 1 (SGA - 1). Luxembourg: Eurostat.

Runci, M.C., G. Di Bella, and F. Cuppone. 2017. "Integrated Education Microdata to Support Statistics Production". In Lauro, N.C., E. Amaturo, M.G. Grassia, B. Aragona, and M. Marino (*Eds.*). *Data Science and Social Research: Epistemology, Methods, Technology and Applications*: 283-290. New York, NY, U.S.: Springer, *Studies in Classification, Data Analysis, and Knowledge Organization*.

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. Roma, Italia: Istat. http://www.istat.it/it/archivio/193056.

Schnetzer, M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk, and M. Moser. 2015. "Quality Assessment of Imputations in Administrative Data". *Journal of Official Statistics - JOS*, Volume 31, Issue 2: 231-247.

Shafer, G. 1992. "Dempster-Shafer Theory". In S.C. Shapiro *(Editor in Chief). Encyclopedia of artificial intelligence*: 330-331. New York, NY, U.S.: Wiley - Interscience.

United Nations Economic Commission for Europe - UNECE. 2021. *Generic Statistical Information Model - GSIM*. Geneva, Switzerland: UNECE. https://unece.org/statistics/modernstats/gsim.

United Nations Economic Commission for Europe - UNECE. 2020. "Linking GSBPM and GSIM". *Draft report*. Geneva, Switzerland: UNECE.

United Nations Economic Commission for Europe - UNECE. 2019. *Generic Statistical Business Process Model - GSBPM. Version 5.1*. Geneva, Switzerland: UNECE. https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51.

United Nations Economic Commission for Europe - UNECE. 2007. *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics*. Geneva, Switzerland: United Nations.

Venturi, M., and G. Di Bella. 2017. "Supporting the use of administrative data in official statistics. The System of Integrated Microdata – SIM and the Quality Report Card of Administrative data – QRCA". Presentation at the *New Techniques and Technologies for Statistics - NTTS 2017*. Brussels, Belgium 13 - 17 March 2017.

Wallgren, A., and B. Wallgren. 2014. *Register-based Statistics: Statistical Methods for Administrative Data (Second edition)*. Hoboken, NJ, U.S.: Wiley, *Series in Survey Methodology*.

Zabala, F., G. Reid, J. Gudgeon, and M. Feyen. 2013. "Quality Measures for Statistical Outputs using Administrative Data". *Statistical Methods*. Wellington, New Zealand: Statistics New Zealand - Stats NZ.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*, Volume 66, Issue 1: 41-63.

Informazioni per le autrici e per gli autori

La collana è aperta alle autrici e agli autori dell'Istat e del Sistema statistico nazionale e ad altri studiosi che abbiano partecipato ad attività promosse dall'Istat, dal Sistan, da altri Enti di ricerca e dalle Università (convegni, seminari, gruppi di lavoro, etc.).

Coloro che desiderano pubblicare su questa collana devono sottoporre il proprio contributo al Comitato di redazione degli Istat working papers, inviandolo per posta elettronica all'indirizzo: iwp@istat.it.

Il saggio deve essere redatto seguendo gli standard editoriali previsti (disponibili sul sito dell'Istat), corredato di un sommario in Italiano e in Inglese e accompagnato da una dichiarazione di paternità dell'opera.

Per le autrici e gli autori dell'Istat, la sottomissione dei lavori deve essere accompagnata da un'e-mail della/del propria/o referente (Direttora/e, Responsabile di Servizio, etc.), che ne assicura la presa visione.

Per le autrici e gli autori degli altri Enti del Sistan la trasmissione avviene attraverso la/il responsabile dell'Ufficio di statistica, che ne prende visione. Per tutte le altre autrici e gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione.

Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del Chicago Manual of Style.

Attraverso il Comitato di redazione, tutti i lavori saranno sottoposti a un processo di valutazione doppio e anonimo che determinerà la significatività del lavoro per il progresso dell'attività statistica istituzionale.

La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line gratuitamente.

Gli articoli pubblicati impegnano esclusivamente le autrici e gli autori e le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.