istat
working
papers

N.18
2016

Il sistema di integrazione dei dati amministrativi in Istat

Maria Carla Runci, Grazia Di Bella, Lorena Galiè

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.18
2016

Il sistema di integrazione dei dati amministrativi in Istat

Maria Carla Runci, Grazia Di Bella, Lorena Galiè

Istat

Istat Working Papers

Il sistema di integrazione dei dati amministrativi in Istat

N. 18/2016

ISBN 978-88-458-1911-7

Il sistema di integrazione dei dati amministrativi in Istat[1]

Maria Carla Runci, Grazia Di Bella, Lorena Galiè

Sommario

Con l'obiettivo di rafforzare l'utilizzo dei dati amministrativi a fini statistici e ottimizzare il processo di produzione, l'Istat si è dotato di una infrastruttura centralizzata per la loro gestione. Il processo di integrazione dei dati prevede come primo passo il caricamento e riconoscimento dei dati amministrativi e l'applicazione successiva di specifiche strategie di record linkage definite sulla base dei domini dell'informazione amministrativa e dipendenti dalla qualità delle variabili identificative delle unità di base.

Parole chiave: elenco parole chiave.

Abstract

Istat has set up a centralised infrastructure for the integration of administrative data in order to enhance their use for statistical purposes and to optimise the statistics production process. The process of data integration is carried out first by uploading and identifying administrative data, then by applying specific record linkage strategies tailored on the administrative information domains and depending on the quality of the identifying variables of the basic units.

Keywords: administrative sources, data integration.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Indice

1. Premessa

La gestione centralizzata dei dati amministrativi costituisce un passo importante dell'istituto verso la modernizzazione dei processi di produzione nell'ottica di massimizzare l'efficienza della produzione e il contenuto informativo delle statistiche.

Il *repository* dei dati amministrativi acquisiti dall'Istat, denominato Sistema Integrato di Microdati (SIM), ha lo scopo di supportare trasversalmente i processi di produzione dell'istituto. I dati vengono considerati micro poiché si riferiscono alle unità: Individui, Unità economiche, Luoghi, ovvero le unità di base della statistica ufficiale. L'aggettivo integrato si riferisce al processo di integrazione ed in particolare all'integrazione delle unità di base. Oltre alle unità di base, che allo stato attuale comprendono le tre tipologie di Individui, Unità economiche e Luoghi, anche le relazioni tra le unità dello stesso tipo o di diverso tipo costituiscono uno specifico interesse per l'analisi dei fenomeni statistici. Le relazioni tra unità dello stesso tipo determinano 'unità complesse' (famiglie, gruppi di imprese, aree di luoghi contigui,…); le relazioni tra unità di tipo diverso danno luogo a sistemi di microdati che permettono la creazione di informazioni integrate molto dettagliate: sistemi di tipo Leed - *Linked employer employee data* (Individui-Unità economiche), sistemi Studenti-Scuole/Atenei (Individui - Unità economiche), sistemi di mobilità sul territorio (Individui - Luoghi). In generale, gli insiemi delle unità di base e le loro relazioni costituiscono l'input per la determinazione del Sistema dei registri: registri delle popolazioni obiettivo e registri delle attività. L'obiettivo è di sfruttare al massimo l'utilizzo dei dati amministrativi migliorando l'efficienza della produzione e il potere informativo delle statistiche.

Nel presente documento si riporta una descrizione del Sistema di integrazione dei dati amministrativi in Istat.

2. L'infrastruttura centralizzata dei dati amministrativi: l'integrazione dei dati

La fonte amministrativa per sua definizione viene generata nell'ambito dell'applicazione di uno o più regolamenti amministrativi e viene gestita da un titolare. Laddove i dati della fonte siano considerati utilizzabili a fini statistici, a seguito di accordi specifici l'Istat procede a formalizzare una richiesta attraverso un modulo ufficiale[2] che comprende:

- un tracciato record (o più tracciati record) che descrive (descrivono) in termini di metadati la selezione/estrazione dalla Fonte amministrativa;
- il riferimento temporale dei dati richiesti;
- una data utile di ricevimento della fornitura (periodo compreso fra una data minima e una data massima concordata con il fornitore).

Il sottoinsieme dei dati amministrativi estratto dalla fonte in base alle specifiche richieste è denominato *Dataset amministrativo* e, più precisamente, è definito nel seguente modo.

Definizione di Dataset amministrativo - Administrative dataset

Un insieme strutturato di dati estratti da una o più fonti amministrative, prima di qualsiasi trattamento o validazione statistica (in ambito Sistan).[3] [1]

[2] Deliberazione 20 aprile 2004, n. 9, Criteri e modalità per la comunicazione dei dati personali nell'ambito del Sistema statistico nazionale, Direttiva n. 9/Comstat, Gazzetta Ufficiale 23 dicembre 2004, n. 300.

[3] Dal Glossario in uso in Istat derivante dall'adattamento del Glossario prodotto nell'ambito del progetto internazionale *ESSnet AdminData,* "Glossary of the main terms related to the use of admin data in producing business statistical", 2012, *https://ec.europa.eu/eurostat/cros/content/glossary-main-terms-related-use-admin-data-producing-business-statistics-first-release_en*

Gli oggetti amministrativi, unità ed eventi, e le variabili amministrative che entrano nel Sistema sono quelli contenuti nei dataset amministrativi.

Il termine *Archivio amministrativo*, utilizzato nei metadati del SIM comprende l'insieme dei dataset amministrativi acquisiti nel tempo da una specifica Fonte. Ad esempio l'*Archivio 115*, *MIUR – Archivio degli iscritti e delle iscrizioni universitarie*, estratto dalla *Fonte amministrativa Anagrafe Nazionale degli Studenti* (ANS) del Miur, comprende tutte le forniture dei corrispondenti dataset (*T*) per ogni *T*.

Da una stessa Fonte l'Istat può acquisire più dataset che possono avere tempi di estrazione diversi, ad esempio dall'ANS, oltre all'*Archivio degli iscritti e delle iscrizioni universitarie* dell'Anno Accademico (*T - T+1*), viene anche acquisito l'*Archivio dei Laureati nell'anno solare T*. Invece, nel caso di acquisizione di dati amministrativi provvisori e definitivi, i due dataset hanno la stessa composizione e lo stesso riferimento temporale, ma vengono estratti dalla Fonte, e successivamente acquisiti, in due momenti diversi.

Di seguito, per comodità, il dataset relativo all'*Archivio r-esimo* con anno di riferimento *T* sarà indicato genericamente come dataset *i-esimo*, con $i = 1, 2, ..., I_t$, dove I_t è il numero totale di dataset amministrativi acquisiti fino al tempo *t*.

Quindi, ogni dataset *i-esimo* è caratterizzato dall'Archivio di riferimento e dalla relativa Fonte amministrativa, dal riferimento temporale e dalla data di estrazione/acquisizione.

Il processo di integrazione comincia con l'operazione di caricamento del dataset. Tale operazione determina il riconoscimento degli oggetti amministrativi presenti nel dataset *i* come di seguito descritto.

Il ciclo di vita del dato amministrativo prevede una prima fase in cui gli oggetti della fonte amministrativa vengono acquisiti in SIM come input, una seconda fase relativa al processo di integrazione e una terza fase in cui l'output di SIM viene messo a disposizione degli utenti interni dell'Istat e utilizzato come input all'interno dei processi di produzione dove prosegue il suo percorso. Il processo centralizzato di gestione dei dati amministrativi ha una funzione di tipo trasversale rispetto agli utenti Istat interni e, attraverso l'utilizzo di codici identificativi univoci permette anche di gestire più agevolmente il rispetto della normativa sulla sicurezza dei dati.

2.1 L'operazione di caricamento e riconoscimento dei dati amministrativi

Il caricamento del dataset amministrativo in SIM consiste nella derivazione di informazioni finalizzate a rendere omogenei tra di loro e aderenti alla logica applicativa del Sistema d'integrazione i dati provenienti da fonti diverse e aventi caratteristiche differenti. A questo scopo viene effettuata un'analisi concettuale dei dati in base allo schema Entità/Relazioni che permette di caricare i dati in tabelle relazionali associando a ciascun oggetto/entità i propri attributi.

Gli oggetti/entità presenti nel dataset vengono riconosciuti come unità statistiche di base di tipo *k* con $k = 1, 2, 3$ corrispondente alle tre tipologie: Individui, Unità economiche e Luoghi. In ciascun dataset possono essere presenti più tipi di unità di base.

Le regole standard per il caricamento nel Data base Oracle di SIM sono le seguenti:

1. Ogni archivio è individuato da un codice che lo identifica all'interno del Sistema; tale codice è memorizzato nei metadati;
2. Gli archivi sono organizzati in una o più strutture dati (tabelle) definite sulla base degli oggetti/entità (unità di analisi/unità di base) e delle relazioni che in essi sono rappresentate.
3. Ogni istanza (riga) di una struttura dati è individuata da: [1]
   - un numero progressivo;
   - uno o più riferimenti temporali (data inserimento, data ultima variazione, anno di riferimento delle informazioni, ecc.);
   - un codice identificativo univoco assegnato alle unità che ne permette l'identificazione all'interno del Sistema;
   - gli attributi dell'oggetto provenienti dalla fonte.

2.2 Inquadramento del processo di integrazione dei dati amministrativi e descrizione delle strutture dei dati e dei metadati

Il processo di integrazione può essere considerato come un processo di identificazione /riconoscimento delle unità del dataset rispetto alle unità già presenti nel Sistema al momento dell'integrazione *t*.

Il processo è di tipo incrementale: via via che i dataset arrivano in Istat vengono progressivamente integrati con i dati già presenti. Operativamente l'integrazione del dataset *i-esimo* avviene attraverso una serie di procedimenti di *record linkage* tra il dataset in ingresso e le cosiddette Basi per l'Integrazione. L'integrazione di un dataset prevede tanti processi di integrazione quanti sono i tipi di unità presenti. Le unità e le relazioni presenti in ciascun dataset vengono coerentemente integrate nell'ambito della stessa tipologia e ciascun elemento entra a far parte della corrispondente Base per l'integrazione B_t^k di SIM, definita nel seguente modo.

Definizione di Base per l'Integrazione B_t^k
Struttura di dati di tipo incrementale, dipendente dalla tipologia di sottosistema d'integrazione, che svolge le seguenti funzioni:

- conserva i legami tra gli elementi del sottosistema di integrazione e i singoli archivi di input;
- costituisce la base informativa rispetto a cui si linkano le unità di tipo *k* dei singoli archivi amministrativi al tempo *t*; [2]
- consente la conservazione nel tempo dell'identificativo dei singoli elementi appartenenti a uno specifico sottosistema di integrazione;
- consente la conservazione nel tempo dell'informazione relativa alla composizione degli archivi di input in cui un determinato elemento appartenente ad un sottosistema d'integrazione è presente.

La tabella B_t^k del db Oracle è determinata, quindi, dall'integrazione progressiva delle unità di base di tipo *k* derivate dai dataset amministrativi caricati in SIM fino al tempo *t*.

L'integrazione delle unità di base termina con l'attribuzione del codice identificativo univoco dell'elemento all'interno della rispettiva Base per l'integrazione.

Tutte le unità identificate nei dataset, ovvero la cui qualità della chiave è sufficiente da permetterne l'identificazione, vengono caricate nelle B_t^k con un codice univoco. Le unità che le procedure di integrazione non abbinano con altre già presenti in SIM avranno assegnato un codice nuovo.

Scopo principale delle Basi per l'integrazione è garantire, quindi, la conservazione nel tempo del codice d'identificazione di una determinata unità elementare all'interno del sottosistema integrato e di mantenere i collegamenti, definiti attraverso il processo di linkage, tra le diverse fonti.

Occorre sottolineare che il grado di integrazione del dataset *i-esimo* (tasso di abbinamento) dipende dal momento in cui avviene l'integrazione ovvero dalla sequenza temporale di caricamento dei dataset man mano che arrivano in Istat: unità che inizialmente non trovano un abbinamento poiché sono presenti per la prima volta nella Base, possono successivamente abbinarsi con unità che progressivamente popolano il Sistema. In questo contesto riveste un ruolo importante anche la tempestività dei dataset ovvero la differenza tra il riferimento temporale dei dati e la data di acquisizione/caricamento.

Ovviamente ci sono dataset che apportano teoricamente nuove unità come, ad esempio, l'Anagrafe Tributaria delle Persone Fisiche, il cui ingresso nel Sistema è determinante per la Base per l'integrazione degli Individui. Questa caratteristica verrà ripresa nel paragrafo 2.2.1. La struttura della B_t^k, come detto [2], pur seguendo una logica comune, dipende dal particolare sottosistema di integrazione a cui si riferisce. Da questo momento in poi la trattazione riguarderà la Base Individui poiché è ritenuta di maggior interesse sia rispetto alle Unità economiche che hanno una storia più consolidata, legata alla costruzione del Registro delle imprese e al relativo Regolamento europeo 696/93, sia rispetto ai Luoghi che presentano delle peculiarità specifiche in relazione al processo di integrazione.

I campi della B_t^k per $k = 1$ (unità di base Individui) sono i seguenti:

ANNO
CODICE_ARCHIVIO
PROGRESSIVO
CODICE_INDIVIDUO
CODICE_FISCALE
COGNOME
NOME
SESSO
DATA_NASCITA
CODCAT_NASCITA
PROV_NASCITA
COM_NASCITA
PAESE_NASCITA
DATA_INGRESSO
DATA_DECESSO
A2009
A2010
A2011
…
A2018
STEP
SUBARCHIVIO

[3]

La variabile ANNO indica l'ultimo anno di inserimento dell'istanza nella Base. Il CODICE_ARCHIVIO individua la fornitura acquisita da una fonte amministrativa e, insieme al riferimento temporale espresso dalle variabili indicatrici di presenza dell'Individuo nell'anno (A2009, A2010,…, A2018), individua univocamente il dataset amministrativo di input. Il PROGRESSIVO, in gerarchia con l'individuazione del dataset amministrativo, permette il collegamento con gli altri dati in esso contenuti (altri attributi specifici delle unità). Il CODICE_INDIVIDUO è il codice univoco che identifica l'Individuo nel tempo all'interno del Sistema. I seguenti campi costituiscono le variabili di linkage per la procedura di integrazione nella Base Individui.

CODICE_FISCALE
COGNOME
NOME
SESSO
DATA_NASCITA
PROV_NASCITA
COM_NASCITA
PAESE_NASCITA

[4]

L'insieme di queste variabili può essere definito come *Rappresentazione dell'unità j-esima nel dataset i* e indicato con il vettore $\underline{x_j}$. Le istanze delle tabelle Individui provenienti dai diversi archivi di input, tabelle di origine descritte in [1], entrano nella B_t^1 con le loro informazioni originali ad eccezione del SESSO, che è sempre codificato (M, F), del CODICE FISCALE, che è impostato a NULL se il valore originale non rispetta definite regole di correttezza formale, del COGNOME e NOME, che sono impostati a NULL se la loro lunghezza (dopo avere eliminato i caratteri speciali) risulta essere = 0, di PROV_NASCITA e COM_NASCITA, che sono ricodificati secondo la codifica Istat, se negli archivi di input l'informazione è disponibile in forma diversa (denominazione, altra codifica).

La variabile STEP indica il passo a cui l'unità si è integrata nella Base, nell'ambito della strategia di integrazione adottata per il dataset. Infine, il campo SUBARCHIVIO permette di distinguere le unità che, pur provenendo dallo stesso archivio riferito all'anno *T*, sono caricati in più tabelle. È il caso di archivi che prevedono più forniture nel corso dell'anno (ad esempio l'Archivio INPS delle denunce retributive mensili - EMens che viene fornito anticipatamente in versione provvisoria e successivamente definitiva) o di archivi che comprendono più tipologie di popolazioni amministrative (ad esempio il 770 statistico che comprende la popolazione dei Dichiarati e la popolazione dei

Sostituti d'imposta con codice fiscale alfanumerico). La variabile SUBARCHIVIO assume un valore progressivo che le distingue.

Quindi, ciascun elemento della popolazione di tipo *k* riferita all'unità di base Individuo è presente in un *grappolo* definito dal Codice Individuo in cui confluiscono tutte le *Rappresentazioni* $\underline{x}_j$ (descritte in [4]) che si sono integrate nel tempo con altre Rappresentazioni già presenti. Nel caso di Individui non riconosciuti come già presenti nella Base, il grappolo contiene ovviamente una sola Rappresentazione.

Alla luce delle recenti nuove frontiere della ricerca in tema di linkage (Steorts et al., 2014) si osserva che le Rappresentazioni di un Individuo confluite in un grappolo e provenienti dai dataset che compongono il SIM suggeriscono pienamente il concetto di «individuo latente».

Nella Base Individui è possibile individuare la popolazione degli Individui riferita all'anno *T* effettuando un filtro per AT = 1 e considerando i record distinti. Tale popolazione contiene tutti gli Individui identificati negli archivi riferiti all'anno *T,* integrati nel Sistema.

Occorre sottolineare che l'anno di riferimento *T* non ha una definizione omogenea nei vari archivi. Ad esempio, per l'Archivio 116 Miur – Dati sul personale docente e non docente universitario, i dati si riferiscono al personale in servizio al *31/12/T,*[4] per l'Archivio 45 Miur – Anagrafe degli Studenti delle scuole i dati si riferiscono all'anno scolastico *T – T+1*, per l'Archivio 9 INPS – Archivio delle denunce retributive Emens, sono presenti tutte le denunce mensili riferite all'anno *T*, ovvero i lavoratori dipendenti dichiarati nell'anno *T*.

2.2.1 Dominio dell'informazione amministrativa

Il processo di integrazione dei dataset nella Base Individui dipende dal dominio dei dati amministrativi dell'archivio, ovvero dalle caratteristiche rispetto alla struttura concettuale e dalle variabili identificative delle unità in esso presenti.

I criteri di classificazione sono:

a) La presenza e la qualità delle variabili di integrazione (grado di linkabilità).
b) L'organizzazione concettuale della entità 'Individuo' rispetto al tempo.
c) L'esistenza di un'unica variabile di identificazione esterna alle variabili dell'archivio per le istanze dell'entità 'Individuo'. [5]
d) La stabilità nel tempo del valore della variabile di identificazione esterna ed i criteri di stabilità;
e) La possibilità che le unità dell'archivio possano generalmente essere già contenute nella Base.[5]

Questi elementi consentono di classificare ciascun archivio, applicare delle opportune strategie di integrazione e definire degli opportuni indicatori di qualità.

Le modalità associate ai criteri sono articolate nel seguente modo.

a) La presenza e la qualità delle variabili di integrazione:
   1 – Esiste solo il codice fiscale come variabile di integrazione;
   2 – Esistono il codice fiscale ed altre variabili;
   3 – Esistono solo altre variabili.

b) L'organizzazione concettuale della entità 'Individuo' rispetto al tempo:
   1 – L'istanza è presente solo una volta indipendentemente dal tempo;

4 Per gli anni di riferimento T precedenti al 2014.
5 Si esclude la possibilità che il dataset e la Base contengano esattamente gli stessi individui.

2 – L'istanza è presente solo una volta nell'ambito dei dati relativi ad un anno di riferimento ma può essere presente più volte nei diversi anni;
3 – L'istanza può essere presente più volte rispetto ad un determinato anno di riferimento.

c) L'esistenza di un'unica variabile di identificazione esterna alle variabili dell'archivio per le istanze dell'entità 'Individuo':
1 – Sì esiste;
2 – No non esiste.

d) La stabilità nel tempo del valore della variabile di identificazione esterna ed i criteri di stabilità:
1 – Sì è stabile sulla base delle variabili utilizzate per l'integrazione o di un sottoinsieme di esse;
2 – Sì è stabile ma sulla base di variabili diverse da quelle usate per l'integrazione;
3 – No non è stabile.

e) La possibilità che le unità dell'archivio possano generalmente essere già contenute nel Sistema:
1 – La popolazione amministrativa non è generata da eventi di ingresso di tipo diretto (nascite, immigrazioni);
2 – La popolazione amministrativa è generata anche da eventi di ingresso di tipo diretto (nascite, immigrazioni).

Alcune di queste combinazioni non sono ammissibili per l'integrazione che presuppone che ogni istanza proveniente da un determinato archivio, all'interno della Base, sia individuata da almeno una variabile (progressivo o codice fiscale).

2.2.2 Le strategie di integrazione dei dati amministrativi

Il processo d'integrazione fa riferimento ai procedimenti di linkage veri e propri e applica le regole d'integrazione.

Il generale problema di *record linkage* è considerato come un problema di classificazione di tutte le coppie di unità generate dal prodotto cartesiano dei due insiemi da integrare in due insiemi disgiunti: abbinati e non abbinati.

Dovendo integrare un dataset nella Base occorre considerare tutte le coppie derivanti dal prodotto cartesiano delle unità del dataset (identificate attraverso le variabili di linkage disponibili) con tutte le Rappresentazioni delle unità presenti nella Base al momento dell'integrazione *t*:

$$n_i \otimes \sum_j m_j^t \qquad [3]$$

dove n_i è il numero di istanze nella tabella Individui del dataset *i-esimo* (a sua volta relativo all'archivio *r* dell'anno *T*) e m_j^t è il numero di Rappresentazioni dell'unità *j-esima* presente nella Base al momento dell'integrazione *t*.

L'integrazione di un dataset opera per passi che si articolano in relazione allo specifico dominio dell'informazione amministrativa del dataset da integrare. Ciascun archivio ha un suo algoritmo di integrazione, definito in base alle tre principali caratteristiche:

- regole che stabiliscono la strategia d'integrazione: l'eventuale suddivisione del processo in più passi, l'insieme di variabili da utilizzare ad ogni passo, le eventuali variabili di blocco eccetera;
- gli algoritmi di riconoscimento da applicare;
- le regole di accettabilità dei collegamenti creati.

Per ciascuna Rappresentazione presente nella Base, la variabile STEP documenta il passo al

quale l'unità si è integrata per la prima volta nella Base. In particolare sono previsti più passi il cui significato dipende dalla disponibilità delle variabili della chiave nel dataset: in genere al primo passo si tenta il linkage per uguaglianza/similitudine di tutte le variabili della chiave disponibili e alle unità che si linkano viene associata la variabile STEP = 1, nei passi successivi le variabili disponibili vengono utilizzate in modo alternato fino ad arrivare ai passi finali in cui viene utilizzato il solo CODICE_FISCALE come variabile di linkage. Dopo aver eseguito tutti i passi, alle unità che non si linkano con altre Rappresentazioni presenti nella Base, viene assegnato un nuovo Codice Individuo e la variabile STEP è posta uguale a 99. Per ogni dataset integrato, l'algoritmo di integrazione è documentato in SIM.

Più dettagliatamente, il processo d'integrazione alimenta le Basi per l'Integrazione nel seguente modo.

Quando l'archivio è già presente in SIM per almeno un'annualità T - h $(h = 1, 2, \ldots)$ si effettua un primo linkage di tipo longitudinale utilizzando una chiave costante nel tempo (codice interno o codice fiscale) che determina l'attribuzione di un progressivo a ciascuna istanza.

La funzione consiste nel ricercare nella Base l'istanza individuata da:

1. lo stesso codice archivio
2. lo stesso identificativo

Qualora esistano altre variabili di accoppiamento oltre il codice fiscale, si ricerca, mediante l'uguaglianza/similitudine di queste variabili, la conferma dell'accoppiamento. Se l'abbinamento è confermato si considerano due casi.

Se la Rappresentazione della B_t^1 a cui l'istanza si abbina presenta la variabile STEP = 1, allora si procede a modificare il relativo record della Base nel modo seguente: si aggiorna la colonna ANNO con l'anno di riferimento dei dati T e si valorizza la colonna AT = 1, aggiornando in questo modo la Rappresentazione dell'unità al tempo T.

L'aggiornamento della Rappresentazione persistente all'interno di uno stesso archivio avviene per motivi di ottimizzazione delle prestazioni e per evitare che nella Base vengano inserite ad ogni processo di linkage un numero di righe eccessivamente elevato. In questo caso la variabile STEP non viene aggiornata e indica, quindi, la strategia di integrazione di primo ingresso.

Se invece, la Rappresentazione a cui si abbina presenta la variabile STEP $\neq 1$ si procede al confronto delle Rappresentazioni del dataset con tutte le Rappresentazioni presenti nella Base[6].

Per le altre Rappresentazioni del dataset riconosciute come nuove (ovvero non presenti in annualità precedenti dell'archivio) e, in generale, per tutte le unità provenienti da archivi caricati nel Sistema per la prima volta, si effettua il linkage con la Base B_t^1, comprese le stesse Rappresentazioni del dataset (in questo modo è possibile individuare eventuali duplicati presenti nel dataset).

Alla fine del processo, se la Rappresentazione $\underline{x_j}$ del dataset in fase di integrazione si accoppia con una Rappresentazione già presente nella Base, si inserisce un nuovo record avente: ANNO = T, CODICE_ARCHIVIO = r, il PROGRESSIVO già inserito nel dataset che permette il ricongiungimento alle altre variabili in esso presenti, le relative variabili di linkage, il CODICE_INDIVIDUO associato alla Rappresentazione $\underline{x_j}^*$ con la quale si è abbinato in SIM, AT = 1 e STEP corrispondente al passo della strategia di integrazione adottata in relazione alle variabili disponibili. Quindi il grappolo caratterizzato da questo CODICE_INDIVIDUO si alimenta di un ulteriore record.

Nel caso in cui al termine del processo di integrazione la Rappresentazione $\underline{x_j}$ non si sia linkata con altre Rappresentazioni già presenti, il relativo record viene inserito nella Base assegnando un nuovo Codice Individuo e ponendo la variabile STEP uguale a 99. La variabile STEP, come già detto, non si aggiorna e definisce il livello a cui avviene la prima integrazione della Rappresentazione con la Base.

[6] Per le unità riconosciute come persistenti, la condizione aggiuntiva di avere STEP = 1 è stata introdotta nel corso del 2015 per garantire la continuità longitudinale in presenza di miglioramenti della qualità della chiave nell'ambito di un archivio (ad esempio miglioramento della completezza del campo CODICE_FISCALE).

Si noti che quando un Individuo all'interno della Base ha più di una Rappresentazione (record del grappolo) e la procedura di integrazione associa una unità del dataset ad una delle Rappresentazioni del grappolo, per la proprietà transitiva, tale unità si associa anche a tutte le altre Rappresentazioni dello stesso grappolo ritenute ad esse equivalenti per coppie.

3. Sviluppi futuri

Il processo di integrazione dei dati amministrativi con i dati da indagine in termini di unità è il successivo passo da affrontare. Se il processo di gestione delle Unità economiche osservate nei dati amministrativi e nelle indagini è già strutturato attraverso il Portale delle Imprese[7] , per quanto riguarda gli Individui occorre definire una opportuna infrastruttura. Recentemente è stata avviata nell'ambito della fase di modernizzazione la progettazione del Sistema dei registri che pone questa questione al centro dell'attenzione.

Il tema della qualità dell'output dei Sistemi di integrazione, sia per la valutazione finale della qualità delle statistiche prodotte utilizzando i dati integrati, sia per il miglioramento della qualità del processo di integrazione costituisce l'ulteriore sfida che l'Istat si appresta ad affrontare.

La valutazione della qualità dei risultati del *record linkage* è particolarmente importante e negli ultimi anni, fin dal lavoro di Chambers (2009), la comunità scientifica internazionale ha dedicato molta attenzione allo studio di metodologie statistiche che tengano esplicitamente conto delle procedure di integrazione applicate per la produzione dei dati, attraverso appositi indicatori della qualità del linkage, per ottenere statistiche non distorte e più efficienti.

Gli indicatori di qualità dell'integrazione dovrebbero rappresentare gli errori che possono derivare dal *record linkage*: falsi abbinamenti e mancati abbinamenti. Per quanto riguarda il SIM, l'errore di falso abbinamento si ha quando la procedura di integrazione assegna lo stesso Codice Individuo a due Rappresentazioni che nella realtà identificano Individui diversi. Quindi la Rappresentazione entra erroneamente nel corrispondente grappolo j-esimo. L'errore di mancato abbinamento si verifica, invece, quando la coppia delle Rappresentazioni si riferisce allo stesso Individuo ma la procedura di record *linkage* non è in grado di riconoscere le due Rappresentazioni come equivalenti. Considerando che il SIM si alimenta in modo incrementale nel tempo con dataset che possono fornire via via informazioni aggiuntive, nel processo di integrazione sono previste operazioni di correzioni ex post. È possibile cioè che l'inserimento di nuove Rappresentazioni permetta di correggere link pregressi rivelatisi errati: due grappoli, ritenuti precedentemente riferiti a due Individui diversi, possono essere riconosciuti appartenenti allo stesso Individuo e fatti confluire in un unico grappolo, ovvero si riconosce un errore di mancato abbinamento o, viceversa, un grappolo dapprima associato ad un unico Individuo può essere scisso in due in quanto si riconosce che le relative rappresentazioni sono riferite a due Individui diversi, in questo caso si riconosce un errore di falso abbinamento. Queste correzioni ex post sono previste e portano, quindi, ad una riduzione dell'errore nel tempo.

Attualmente sono in corso alcune attività aventi l'obiettivo di valutare gli errori di abbinamento nel contesto del *linkage* probabilistico[8]. In particolare, al fine di stimare l'errore di falso e mancato abbinamento di un dataset amministrativo, è stata condotta una sperimentazione che ha riguardato l'archivio dei Permessi di soggiorno. La scelta dell'archivio, che comprende gli Individui stranieri che richiedono o rinnovano il loro permesso di soggiorno, è stata dettata dalle particolari problematicità che esso presenta in fase di integrazione dovute alla qualità delle variabili di *linkage*. L'applicazione è stata effettuata su un campione del dataset dei Permessi di soggiorno 2014, selezionato in base all'anno di nascita (anno = 1987) in considerazione del fatto che tale variabile presenta una migliore qualità rispetto alle altre variabili di *linkage*. Analogamente il sottoinsieme della

[7] https://imprese.istat.it

[8] Tiziana Tuoto, nell'ambito del Gruppo di Lavoro Istat denominato Arcolaio (Valutazione della qualità statistica degli archivi amministrativi, del loro processo in SIM e delle statistiche prodotte a partire da dati amministrativi), ha curato gli aspetti metodologici dell'attività.

Base Individui è stato estratto con la medesima regola dell'anno di nascita, prendendo tutte le Rappresentazioni dei grappoli con almeno un record rispondente a tale criterio. Il record *linkage* probabilistico ha permesso di selezionare un sottoinsieme di coppie da sottoporre a *clerical review*. L'attività è tuttora in corso e i risultati saranno diffusi in una successiva pubblicazione.

Per quanto attiene al miglioramento del processo, sono in esame alcune attività volte a verificare l'applicabilità di un approccio di tipo modulare definito sulla base dei risultati di monitoraggio della qualità. La caratterizzazione dei domini dei dati amministrativi, sopra esposti, insieme ad appositi indicatori di qualità (sia indicatori di qualità dell'input che indicatori di monitoraggio della qualità del *linkage*) possono definire un sistema di *allert* per l'individuazione di elementi di criticità da risolvere attraverso un processo guidato di *clerical review*.

Dalle prime evidenze emerge che, in generale, la qualità dell'integrazione per le Unità economiche, ma anche per gli Individui è molto buona. La misurazione e la riduzione dell'errore è il prossimo obiettivo.

Continuano, in ogni caso, le attività di collaborazione con gli enti titolari delle fonti amministrative al fine di creare quel necessario coinvolgimento volto al progressivo miglioramento della qualità dei dati in termini di 'linkabilità' delle unità nel rispetto della normativa sulla riservatezza.

Riferimenti bibliografici

Ambroselli S. 2015. *I codici identificativi univoci all'interno del SIM (Sistema Integrato di Micro-dati)*, Istat Working papers, n. 5, Roma.

Chambers, R. 2009. *Regression Analysis Of Probability-Linked Data*. Official Statistics Research Series 4.

ESSnet AdminData. 2013. *Admin Data Glossary. Definitions adopted for certain terms related to the use of administrative data for producing business statistics*, http://www.cros-portal.eu/content/final-release-glossary.

Scanu, M. 2003. *Metodi statistici per il record linkage*. Istat, Collana Metodi e Norme, n. 13.

Steorts, R., Hall, R., Fienberg S.E. 2015. "A Bayesian Approach to Graphical Record Linkage and De-duplication". In *Journal of the American Statistical Association: Theory and Methods* (2015).

Wallgren A. and Wallgren B. 2007. *Register-Based Statistics. Administrative Data for Statistical Purposes*, Wiley.

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