

ARMONIZZARE I MICRODATI DI ARCHIMEDE NEL SISTEMA INTEGRATO DEI REGISTRI

UNO STUDIO DI FATTIBILITÀ





ARMONIZZARE I MICRODATI DI ARCHIMEDE NEL SISTEMA INTEGRATO DEI REGISTRI: UNO STUDIO DI FATTIBILITÀ

EDIZIONE 2021

Attività editoriali: Nadia Mignolli (coordinamento), Marzia Albanesi, Patrizia Balzano e Alessandro Franzò.

Copertina: Maurizio Bonsignori.

ISBN 978-88-458-2059-5

© 2021

Istituto nazionale di statistica
Via Cesare Balbo, 16 - Roma



Salvo diversa indicazione, tutti i contenuti pubblicati sono soggetti alla licenza Creative Commons - Attribuzione - versione 3.0.
<https://creativecommons.org/licenses/by/3.0/it/>

È dunque possibile riprodurre, distribuire, trasmettere e adattare liberamente dati e analisi dell'Istituto nazionale di statistica, anche a scopi commerciali, a condizione che venga citata la fonte.

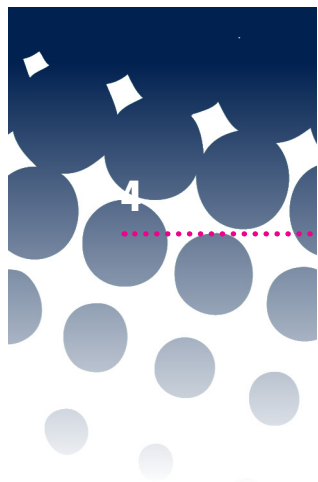
Immagini, loghi (compreso il logo dell'Istat), marchi registrati e altri contenuti di proprietà di terzi appartengono ai rispettivi proprietari e non possono essere riprodotti senza il loro consenso.



INDICE

	Pag.
Introduzione	5
1. Base integrata “Condizioni socio economiche delle famiglie - Archimede”	9
1.1 Premessa	9
1.2 Modellazione del processo di produzione di “Condizioni socio-economiche delle famiglie - Archimede”	9
1.2.1 <i>As-is del processo</i>	9
1.2.2 <i>Le fonti di input e le loro tempistiche</i>	10
1.2.3 <i>I processi</i>	11
1.2.4 <i>Gli attori</i>	11
1.2.5 <i>Definizioni delle unità statistiche di analisi</i>	12
1.2.6 <i>Definizioni principali</i>	12
1.3 Misure di qualità dell’output rispetto alle statistiche ufficiali e indicatori di qualità	13
1.3.1 <i>Coerenza rispetto alla popolazione residente</i>	14
1.3.2 <i>Coerenza rispetto al grado di istruzione</i>	14
1.3.3 <i>Coerenza rispetto al reddito familiare</i>	16
1.3.4 <i>Coerenza rispetto al reddito individuale</i>	17
1.3.5 <i>Coerenza rispetto alla distribuzione delle famiglie</i>	18
1.4 Analisi di corrispondenza tra informazioni del Registro base degli Individui (RBI) e quelle contenute nella base	19
1.4.1 <i>Controlli su macro-dati</i>	20
1.4.2 <i>Controlli su micro-dati</i>	21
1.5 Riflessioni conclusive, punti aperti e criticità	23
2. Base integrata “Popolazione insistente giornaliera per studio e lavoro - Archimede”	25
2.1 Premessa	25
2.2 Informazioni per la redazione dell’analisi as-is del processo	25
2.3 Analisi as-is del processo	26
2.3.1 <i>Schema generale</i>	26
2.3.2 <i>Schema di dettaglio dei sotto-processi (fasi)</i>	27
2.4 Mapping tra contenuti e definizioni della base “Popolazione insistente giornaliera per studio e lavoro - Archimede” e quelle del Registro base degli individui (RBI)	30





Armonizzare i microdati di Archimede nel sistema integrato dei registri: uno studio di fattibilità

2.5 Alcune analisi di qualità: confronti con le statistiche ufficiali	30
2.6 Confronto tra informazioni contenute nella base e nel Registro base degli Individui (RBI) al 31 dicembre 2017	33
2.6.1 <i>Concordanza rispetto alle unità</i>	33
2.6.2 <i>Confronti tra le distribuzioni degli individui a livello territoriale subnazionale</i>	33
2.6.3 <i>Concordanza rispetto alle variabili core del Sistema integrato dei registri (SIR)</i>	34
2.7 Prime conclusioni e sviluppi futuri	35
3. Base integrata “Precarietà lavorativa - Archimede”	37
3.1 Premessa	37
3.2 Contesto	37
3.3 Contenuti informativi della base dati e definizioni principali	38
3.4 Modellazione del processo	38
3.4.1 <i>Mapping tra contenuti della base “Precarietà lavorativa - Archimede” e quelle del Registro base degli Individui (RBI) e definizioni</i>	40
3.5 Qualità rispetto alle statistiche ufficiali	40
3.6 Confronto tra informazioni contenute nella base e nel Registro base degli Individui (RBI)	43
3.7 Conclusioni e sviluppi futuri	44
4. Modellazione del processo di produzione delle basi di Archimede armonizzato con il Sistema integrato dei registri (SIR)	47
4.1 Premessa	47
4.2 Modellazione del processo di produzione delle basi di Archimede armonizzato con il Sistema integrato dei registri (SIR)	47
4.2.1 <i>Modellazione del processo di armonizzazione della Base Condizioni socio economiche delle famiglie - Archimede</i>	48
4.2.2 <i>Modellazione del processo di armonizzazione della Base Popolazione insistente - Archimede</i>	52
4.2.3 <i>Modellazione del processo di armonizzazione della Base Precarietà lavorativa - Archimede</i>	54
4.3 Scenari evolutivi dell’armonizzazione di Archimede con il Sistema integrato dei registri (SIR)	56
4.4 Valutazione di impatto della nuova architettura e proposta di aggiornamento in fasi	59
Riferimenti bibliografici	61
Appendice	63

INTRODUZIONE¹

Nell'ambito del suo processo di modernizzazione avviato nel 2016 l'Istat si è dotato di una infrastruttura che rappresenta il fondamento del nuovo impianto produttivo delle statistiche ufficiali basate sui dati amministrativi: il Sistema integrato dei registri (SIR) (Di Zio *et al.*, 2016). Realizzando la gestione unitaria di registri statistici che si sviluppano su diverse aree tematiche, il SIR ne garantisce l'integrazione concettuale e metodologica. Esso è, infatti, composto da quattro registri principali (Registro base degli individui, delle famiglie e delle convivenze; Registro di base delle unità economiche; Registro degli edifici e del territorio; Registro delle attività) e da alcuni registri satellite (per esempio il Registro dei redditi). Il SIR genera unità statistiche *core* (ad es. Individui residenti, Individui occupati ecc.) e variabili *core* (ad es. Sesso, Età, Grado di istruzione, Cittadinanza, Comune di residenza ecc.), che via via che si renderanno disponibili dovranno essere utilizzati in tutti i processi di produzione dell'Istat al fine di allinearne l'informazione.

Il progetto Archimede (ARCHivi integrati di Microdati Economici e DEMografici) (Istat, 2014) comprende quattro basi integrate di microdati (da qui in poi Basi), tutte iscritte nel Programma statistico nazionale². Tra gli *output* tematici delle Basi ci sono il Comune di lavoro, il Comune di studio, il Reddito familiare lordo, la Condizione di lavoro stabile o non stabile degli occupati regolari (Istat, 2014). Inoltre, poiché il progetto è nato nel 2013, ovvero prima del SIR, tutti gli input delle Basi arrivano dal Sistema integrato dei microdati (SIM) (Ambroselli, 2015) piuttosto che dal SIR.

Gli *output* tematici delle Basi sono utilizzati da utenti interni ed esterni sia in progetti di ricerca, sia in fasi di produzione. Gli enti del Sistan (Sistema statistico nazionale) possono accedervi dal 2015 attraverso la piattaforma Armida (ARchivio dei file di MIcroDati validati) per la realizzazione dei propri progetti iscritti nel Piano statistico nazionale o svolti per finalità istituzionali. Dal 2018 un set di indicatori comunali viene diffuso nell'ambito del portale "A misura di comune" (Istat, 2018). Dal 2019 viene pubblicato un set di tavole di dati per raggruppamenti territoriali funzionali all'interno della Statistica sperimentale "Popolazione insistente per studio e lavoro - Archimede"³ (Istat, 2020).

I dati di *output* di Archimede sono altresì utilizzati nell'ambito di progetti all'interno del Laboratorio innovazione dell'Istat, di gruppi di lavoro, di *task force* e per lo studio di indicatori sul benessere equo e sostenibile. Inoltre, alcune variabili delle Basi sono impiegate come supporto al disegno di campionamento di indagini (per esempio nell'indagine Aspetti della vita quotidiana). Infine, sono utilizzate per il calcolo di stime da modello di alcuni indicatori.

Questo ampio livello di utilizzo rende importante integrare negli attuali processi di produzione di Archimede gli standard metodologici e architetturali e i quadri concettuali di riferimento che l'Istituto sta via via perfezionando, a partire dall'acquisizione delle variabili *core* oggi prodotte direttamente dal SIR, così come faranno tutte le filiere produttive dell'Istat.

¹ La cura del volume e la redazione dell'introduzione sono di Roberta Vivio.

² "Popolazione insistente giornaliera per studio e lavoro - Archimede - IST 02661", "Condizioni socio economiche delle famiglie - Archimede - IST 2663", "Precarietà lavorativa - Archimede - IST 02813", e "Popolazione insistente in aree a rischio - Archimede - IST 02838" (quest'ultimo non oggetto del presente studio di fattibilità).

³ Popolazione insistente per studio e lavoro. Istat. Anno 2018 <https://www.istat.it/it/archivio/258234>. Anno 2017 <https://www.istat.it/it/archivio/248054>; Anno 2016 <https://www.istat.it/it/archivio/239358>.

La maturità attualmente raggiunta dal SIR ha fornito l'opportunità e la necessità di iniziare una riflessione su vantaggi e criticità derivanti dall'armonizzazione di Archimede con il SIR stesso. È con questa finalità che è stata istituita una *task force*, i cui lavori e risultati sono qui raccolti, che ha svolto uno studio di fattibilità sulla produzione armonizzata tra il SIR e le Basi del progetto Archimede, anche al fine di ampliare l'offerta informativa del SIR.

Lo studio effettuato ha avuto una doppia valenza: da un lato si è inteso costituire la base di riferimento per la fase di implementazione vera e propria che sarà svolta in collaborazione dai responsabili dei registri di base e tematici e da quelli dell'attuale processo informativo e tematico di Archimede. Dall'altro, lo studio ha consentito di disegnare il *business case* per analoghi studi di armonizzazione.

Le attività della *task force* hanno riguardato l'armonizzazione rispetto al principale registro del SIR, ovvero il Registro base degli individui, delle famiglie e delle convivenze ed il disegno dei possibili scenari di armonizzazione con le Basi di Archimede. Più in dettaglio sono state realizzate:

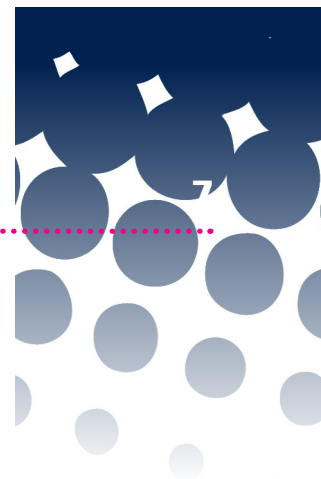
- a. l'analisi *as-is* delle caratteristiche tecniche ed architetture del processo di produzione delle Basi;
- b. l'analisi della corrispondenza tra le informazioni di RBI e quelle contenute nelle Basi al fine di determinare le variabili rispetto alle quali studiare l'armonizzazione;
- c. la misurazione della qualità delle statistiche prodotte dalle Basi rispetto alle statistiche ufficiali di fonte indagine/rilevazione e alle statistiche di fonte RBI;
- d. l'analisi dei requisiti architetture e tecnici e disegno dei possibili scenari di armonizzazione (es. riutilizzo parziale delle procedure in essere; disegno, allo stato attuale e prospettico, del SIR), *gap analysis*;
- e. la proposta di un'architettura *to-be* di integrazione con il SIR, suddivisa in possibili scenari di armonizzazione, sulla base di considerazioni di fattibilità in funzione dello stato di avanzamento del SIR e nell'ottica di riutilizzo delle procedure in essere.

Nei quattro capitoli del volume saranno esposti i risultati delle attività di cui sopra. In particolare, nel Capitolo 1 si esporranno le caratteristiche della base "Condizioni socio-economiche delle famiglie - Archimede", la mappatura delle informazioni, l'analisi di qualità dei dati prodotti rispetto alle statistiche ufficiali di riferimento e rispetto alle variabili comuni con RBI. Un analogo percorso di approfondimento è presentato nel Capitolo 2 per "Popolazione insistente giornaliera per studio e lavoro - Archimede", e nel Capitolo 3 per "Precarietà lavorativa - Archimede".

Il Capitolo 4 è dedicato al nuovo disegno dell'architettura delle basi (versione *to-be*) e a riflessioni sui possibili scenari di evoluzione tra la versione architetture attuale e quella a tendere (*gap analysis*), avendo presente l'attuale stato di avanzamento del SIR e i registri attualmente disponibili. Essi, infatti, sono a diversi livelli di maturazione, per cui non è possibile armonizzare tutto in questa fase di lavoro, ma si può tener conto già in fase di disegno, come si è cercato di fare, delle future acquisizioni informative principali.

Tra i principali punti che verranno sviluppati alcuni sono:

- la coerenza delle Basi pre-armonizzazione rispetto alle statistiche ufficiali relative a variabili *core* e a variabili tematiche raggiunge degli ottimi livelli e garantisce la robustezza delle statistiche già prodotte. I riferimenti sono "Popolazione residente comunale per sesso, anno di nascita e stato civile" relativamente a sesso ed età; "Popolazione residente comunale straniera per sesso ed anno di nascita" relativamente alla cittadinanza; "Rilevazione sulle forze di Lavoro (Rfl)" relativamente a grado di



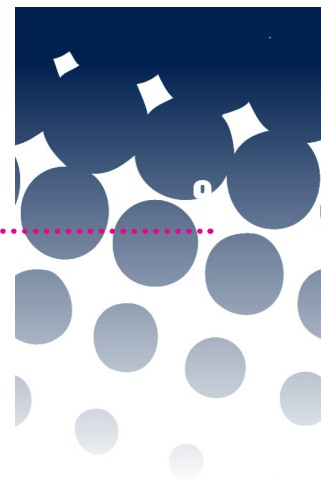
istruzione, condizione professionale e tipologia contrattuale; “Indagine su reddito e condizioni di vita (Eu-Silc)” per il reddito.

- Il *mapping* ha identificato come variabili sostituibili: Sesso, Età, Comune di residenza, Grado di istruzione e Cittadinanza. Anche in questo caso la coerenza tra le Basi e la fonte di confronto (RBI) è molto alta⁴.
- I due precedenti risultati consentono di affermare che l'alimentazione sostitutiva non genererà interruzioni di serie storiche nei dati già prodotti e diffusi da Archimede, e realizzerà un miglioramento in termini di efficienza nel processo di produzione e di armonizzazione concettuale, informativa e architettuale complessiva.
- Dal punto di vista della tempistica, la sola armonizzazione rispetto ad RBI non consentirà alcun miglioramento, dato che le principali fonti di input sono le stesse. Un guadagno di una annualità nel rilascio rispetto all'anno di riferimento dei dati (da due anni ad un anno) potrà realizzarsi in seguito all'utilizzo di altri registri oltre ad RBI (in particolare, il Registro statistico di Base dei Luoghi e il Registro Tematico del Lavoro). L'alimentazione dal SIR, inoltre, consentirebbe una riduzione dei tempi di ulteriori uno-due mesi poiché renderebbe i tre progetti indipendenti su alcuni input (es. Decessi, Popolazione residente).
- Per quanto riguarda l'architettura *as-is* e *to-be*, l'analisi condotta rispetto al meta modello dei Registri di base nel SIR e a quello delle Basi integrate dinamiche del *Register Analytics Framework* (RAF⁵) evidenzia che il metamodello *as-is* delle Basi di Archimede è intermedio rispetto ad essi: è, infatti, una base integrata statica, prodotta accedendo in modo predefinito ed organizzato a fonti del SIM. Alcuni input possono essere sostituiti con alimentazione dal SIR (da RBI, e in prospettiva dagli altri registri), ma per quelli non sostituibili deve restare l'alimentazione da SIM, seppur ridotta rispetto al processo attuale.
- L'utilizzo dei Registri, anche del solo RBI, consente di ridurre la quantità di dati (fonti amministrative) utilizzati per la costruzione delle Basi, e dunque *by design* realizza il principio di minimizzazione del Regolamento generale europeo sulla protezione dei dati personali.
- L'architettura del *to-be* è stata disegnata come una struttura aperta, modulare, e concepita in prima istanza per recepire l'integrazione rispetto a RBI, ma già predisposta ad acquisire successive integrazioni da altri registri del SIR via via che saranno disponibili. In entrambi i casi è aperta anche al SIM.

Dal punto di vista della sicurezza dei dati la predisposizione di due aree di lavoro indipendenti (*working* e *output*) è una delle misure tecniche atte a ridurre i rischi di violazione dei dati personali, e fa parte di un insieme di misure che si stanno predisponendo per conformarsi al regolamento europeo.

⁴ L'alimentazione sostitutiva è stata analizzata rispetto a RBI 2017: in questa annualità non sono implementati i pesi/correttivi atti a riproporzionare i conteggi di popolazione in fase di *output*.

⁵ Nel disegno degli scenari si è tenuto conto delle prerogative e caratterizzazioni del RAF che al momento in cui dispiegherà la sua architettura prevederà dei tool di accesso al SIR per la realizzazione dinamica di Basi integrate. Al momento il RAF è in fase di disegno.



1. BASE INTEGRATA “CONDIZIONI SOCIO ECONOMICHE DELLE FAMIGLIE”¹

1.1 Premessa

Il presente capitolo vuole fornire un report sullo stato di avanzamento dell'iter per la produzione armonizzata tra la base integrata di microdati amministrativi “Condizioni socio economiche delle famiglie - Archimede” ed il Sistema integrato dei registri (SIR). In particolare, illustra i risultati dello studio di fattibilità. Il Paragrafo 1.2 presenta la modellazione – effettuata tramite il linguaggio Archimate² – dell'attuale processo di produzione (*as-is*); nel Paragrafo 1.3 vengono illustrati i risultati relativi alle misure di qualità della base sulle condizioni socio-economiche delle famiglie (pre-armonizzazione) rispetto alle statistiche ufficiali. Infine, nel Paragrafo 1.4 viene descritta l'attività di *mapping* delle informazioni contenute nella base con quelle presenti nei registri – in particolare nel Registro base degli individui, delle famiglie e delle convivenze (RBI) – e sono presentati le principali evidenze emerse dall'analisi di corrispondenza tra le due fonti.

1.2 Modellazione del processo di produzione di “Condizioni socio-economiche delle famiglie - Archimede”

1.2.1 As-is del processo

La rappresentazione (*as-is*) dell'attuale processo di produzione³ è stata realizzata seguendo i principi dell'*Enterprise Architecture* (Lankhorst, 2018), utilizzando il linguaggio Archimate, mediante il quale sono stati esplicitati le fonti acquisite, le strutture Istat coinvolte (direzioni/servizi e relativi ruoli), i principali processi che concorrono alla creazione della base “Condizioni socio economiche delle famiglie - Archimede”, gli output prodotti.

La fase di modellazione è stata preceduta da un'attività di analisi del processo esistente in cui si sono presi in esame, oltre agli elementi appena citati, anche altri aspetti, come le popolazioni di riferimento coinvolte, gli aspetti definitivi, i dati presenti nei tracciati di input ed i relativi trattamenti operati, i vincoli, gli aspetti di criticità da superare, le interconnessioni con altri processi, le tempistiche connesse alle disponibilità delle fonti e la tempestività nella predisposizione della base in oggetto. La fase di modellazione ha portato al disegno dello schema generale del processo di produzione (riportato in Figura 1.1) più altri sette schemi di dettaglio per ciascuna delle sotto-fasi in cui è stato possibile scomporre il processo di produzione esaminato. A sinistra dello schema in Figura 1.1 sono riportate tutte le fonti di input utilizzate per la costruzione della base (raggruppate per area tematica), in alto

¹ I Paragrafi 1.1, 1.2.5, 1.2.6, 1.3 e 1.5 sono a cura di Sara Casacci; i Paragrafi 1.2.1, 1.2.2, 1.2.3, 1.2.4 e l'Appendice sono a cura di Domenico Infante; il Paragrafo 1.4 è a cura di Stefania Di Domenico.

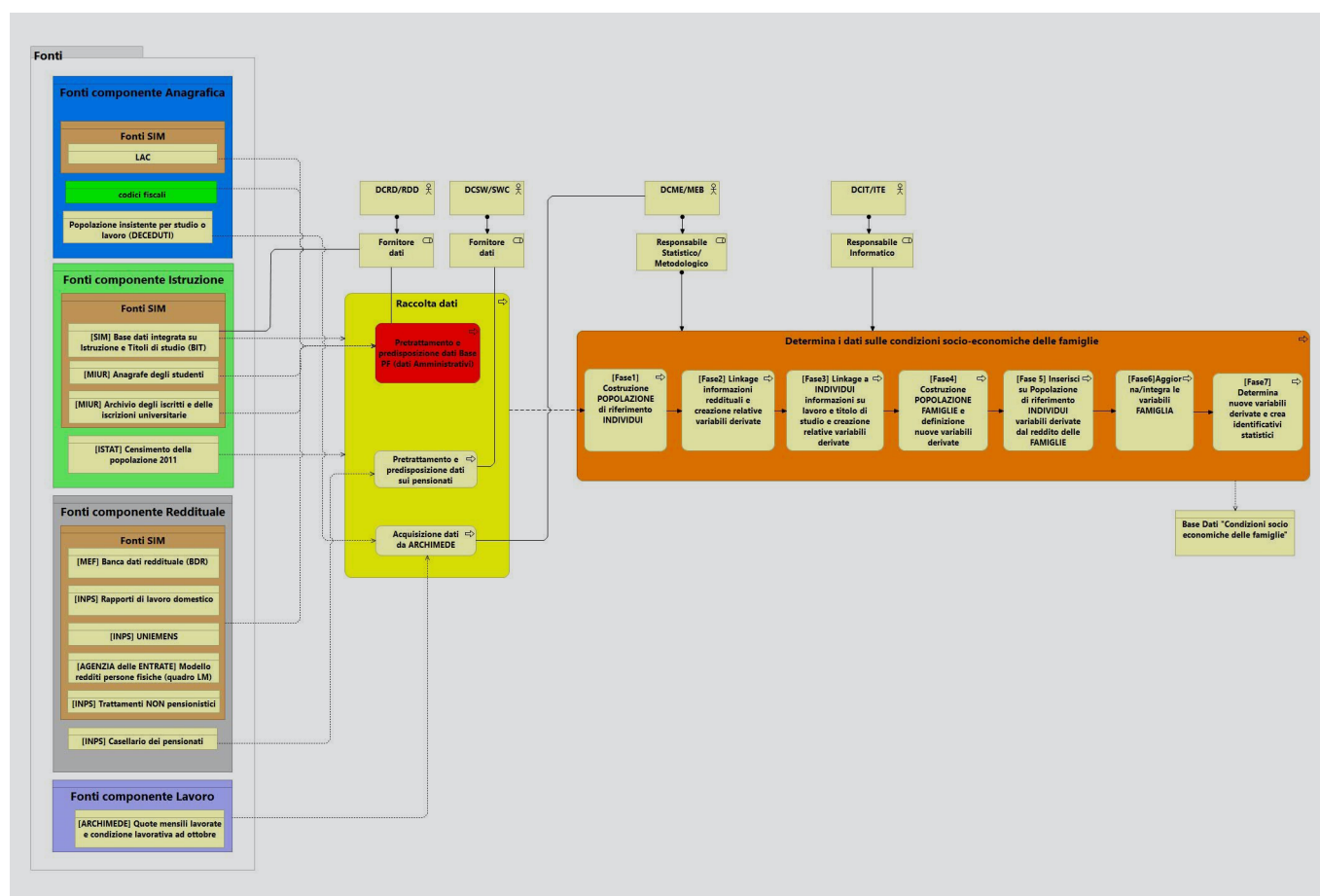
² <https://www.opengroup.org/togaf>.

³ Il periodo di riferimento dei dati è il 2017.



sono riportati i ruoli e gli attori coinvolti (direzione/servizio Istat), nella parte centrale della Figura 1.1 i sotto processi individuati ed in basso a destra il prodotto finale, ossia la base dati “Condizioni socio economiche delle famiglie - Archimede”.

Figura 1.1 - As-is schema generale del processo di produzione relativo alla base dati Condizioni socio-economiche delle famiglie - Archimede



Fonte: Elaborazione degli autori

1.2.2 Le fonti di input e le loro tempistiche

Le fonti in input del processo (Prospetto 1.1) sono costituite da archivi amministrativi e basi integrate statistiche. Gli archivi sono acquisiti dal Servizio Fonti amministrative e integrazione dei registri (RDD), struttura che gestisce il Sistema integrato di microdati (SIM) ed effettua una lavorazione *ad hoc* su alcuni di essi mettendo a disposizione una base informativa preliminare. Le basi statistiche provengono da altri settori di produzione Istat, tra cui lo stesso progetto Archimede.

Nel Prospetto 1.1 è possibile osservare, oltre all'elenco delle fonti utilizzate (periodo di riferimento 2017), anche l'informazione sulle tempistiche di acquisizione in Istat. Tale disponibilità è espressa in termini di data massima della ricezione della fornitura in Istat, come riportata nella *Quality report card* dei dati amministrativi (QRCA).

1. Base integrata “Condizioni socio economiche delle famiglie”

L'acquisizione di alcune fonti, ad esempio quelle di natura fiscale, sconta un lag temporale di T+2 anni.

Prospetto 1.1 - Elenco delle fonti utilizzate per la costruzione della base dati e tempistiche relative alla loro disponibilità

Fonte (ENTE TITOLARE)	DISPONIBILITÀ FONTE (DATA MAX RICEZIONE FORNITURA QRCA)
Liste anagrafiche comunali - LAC (Comuni)	T+3 mesi (Dati riferiti allo 01/01/2020 acquisiti entro il 31/03/2020)
Banca dati statistica reddituale (Mef)	T+2 anni (Dati riferiti al 2018 acquisiti entro il 31/07/2020)
Rapporti di lavoro domestico (Inps)	T+1 anno (Dati riferiti al 2019 acquisiti entro il 30/05/2020)
Prestazioni non pensionistiche (Inps)	- Trattamenti monetari non pensionistici t+1 anno (Dati riferiti al 2019 acquisiti entro il 30/10/2020)
	- Assegni al nucleo familiare t+2 anni (Dati riferiti al 2018 acquisiti entro il 29/02/2020)
Anagrafe nazionale degli studenti delle scuole (Miur)	T+2 anni (Dati riferiti al 2018 acquisiti entro il 31/01/2020)
Anagrafe nazionale degli studenti universitari (Miur)	T+2 anni (Dati riferiti al 2018 acquisiti entro il 31/03/2020)
Modello Redditi Persone Fisiche - quadro LM (Agenzia delle Entrate)	T+1 anno (Dati riferiti al 2019 acquisiti entro il 31/12/2020)
Uniemens – parte contributiva (Inps)	T+1 anno (Dati riferiti al 2019 acquisiti entro il 30/10/2020)
Casellario dei pensionati e dei trattamenti pensionistici (Inps)	T+1 anno (Dati riferiti al 2019 acquisiti entro il 30/09/2020)
Censimento della popolazione 2011 (Istat)	-
Base dati integrata su istruzione e titoli di studio -BIT (Istat)	Fonte non gestita da QRCA, disponibile con un ritardo variabile tra 12 e 24 mesi
Precarietà lavorativa - Archimede (Istat)	
Popolazione insistente - Archimede (Istat)	

Fonte: *Quality Report Card* dei dati amministrativi

1.2.3 I processi

Nello schema generale dell'*as-is* (Figura 1.1) sono rappresentati i due macro-processi principali. Il primo è il processo di “Raccolta dati”, all'interno del quale sono riportati tre sotto-processi, a rappresentare: (1) l'estrazione delle fonti prodotte da SIM, (2) l'estrazione dei dati fornita dalla Direzione Centrale per le statistiche sociali e il *welfare* dell'Istat e (3) le fonti attinte da altre basi di Archimede, realizzate dalla Direzione Centrale per la metodologia e il disegno dei processi statistici dell'Istat.

Il secondo processo principale di “Determina i dati delle condizioni socio-economiche delle famiglie” elabora i dati dalle fonti di input e produce la base; all'interno del blocco, sono rappresentate le sette sotto-fasi in cui è scomponibile questo blocco. Gli schemi di ciascuna sotto-fase con i relativi oggetti di input e di output sono riportati in Appendice.

1.2.4 Gli attori

Come anticipato, gli attori Istat coinvolti nel processo e i loro ruoli sono mappati nella parte alta dello schema (Figura 1.1). In dettaglio, essi sono:

- La Direzione centrale per la metodologia e il disegno dei processi statistici (DCME), che compare sia col ruolo di fornitore di dati di input (per la quota parte di informazioni provenienti da altre basi di Archimede) sia col ruolo di responsabile statistico/metodologico, curando sia la progettazione del processo, sia le fasi di verifica e supervisione delle attività.
- La Direzione centrale per le tecnologie informatiche (DCIT), nel ruolo di responsabile informatico che, oltre a curare la progettazione e la realizzazione delle componenti software, è parte attiva del processo in quanto cura l'esecuzione delle procedure applicative modellate nelle sotto-fasi del processo di “Determina i dati delle condizioni socio-economiche delle famiglie”.
- La Direzione centrale per le statistiche sociali e il *welfare* (DCSW) che ha ruolo di fornitore per una delle fonti di input (fino al periodo di riferimento dei dati 2017).



- La Direzione centrale per la raccolta dati (DCRD) che ha il ruolo di fornitore di dati di input ed è titolare del “Pretrattamento e predisposizione della base preliminare di Precarietà Familiare”.

1.2.5 Definizioni delle unità statistiche di analisi

Le unità di analisi sono gli Individui residenti in famiglia (non sono considerati gli Individui residenti in convivenza) e le Famiglie anagrafiche. La Famiglia anagrafica è definita come “l’insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o affettivi, coabitanti ed aventi dimora abituale nello stesso comune. Può essere costituita anche da una sola persona”.

L’universo di riferimento è costituito dall’insieme delle famiglie i cui componenti risultano iscritti nelle Liste anagrafiche comunali (Lac) alla data di riferimento (1 gennaio 2018 per la base dati del 2017).

1.2.6 Definizioni principali

Lavoratori: sono considerati Lavoratori gli individui con segnali di occupazione regolare negli archivi amministrativi nel mese di ottobre (mutuata dalla base “Precarietà lavorativa - Archimede”).

Intensità lavorativa familiare: rappresenta una misura della partecipazione dei componenti della famiglia al mercato del lavoro (solo occupazione regolare derivata dalle fonti amministrative) nel corso dell’anno. L’informazione viene rilevata a livello individuale e poi aggregata a livello familiare. La quota annuale a livello individuale è computata come media dell’intensità lavorativa mensile (informazione fornita dalla base “Precarietà lavorativa - Archimede”). Quest’ultima viene calcolata per tutti i mesi dell’anno e rappresenta una misura della occupazione mensile del singolo individuo (varia tra zero - nessun giorno di lavoro nel mese - e uno - lavoro continuativo nel mese). Osservate tutte le posizioni lavorative che un soggetto ricopre in un determinato periodo, l’intensità lavorativa mensile complessiva è calcolata sommando le intensità di tutte le posizioni (con somma massima unitaria); a livello annuale, l’intensità lavorativa è il rapporto tra la somma delle intensità mensili e il numero di mesi. Ai fini del calcolo della misura a livello familiare si considerano i membri della famiglia di età compresa tra i 18 e i 59 anni, escludendo gli studenti nella fascia di età tra i 18 e i 24 anni.

Reddito lordo: è il reddito complessivo al lordo della tassazione, ed è calcolato come somma di diverse componenti di natura economica percepite dai membri della famiglia anagrafica nel corso dell’anno. Le principali voci incluse nel reddito lordo sono i redditi da lavoro dipendente e autonomo, da pensione, da capitale, da trasferimento pubblico e privato. La soluzione adottata prevede il recupero di voci di reddito dalle fonti amministrative utilizzate indipendentemente dal fatto che esse siano tassate o esenti e dall’eventuale tipologia di tassazione (ordinaria, separata). Alle variabili reddituali derivanti dall’utilizzo delle fonti corrisponde una definizione di tipo amministrativo permeata su concetti prevalentemente di tipo fiscale e retributivo, associabile a specifiche sotto-popolazioni di contribuenti e/o percettori di un reddito certificato. Per questo motivo la somma di queste componenti, che definisce la variabile di reddito lordo, presenta dei disallineamenti con una definizione statistica di reddito.

Reddito lordo equivalente: è una trasformazione del Reddito lordo che consente di confrontare i redditi di famiglie strutturalmente diverse per numerosità ed età dei componenti effettuata

1. Base integrata “Condizioni socio economiche delle famiglie”

con l'utilizzo di una scala di equivalenza che tiene conto non solo del numero ma anche dell'età dei componenti della famiglia. La scala di equivalenza modificata utilizzata per calcolare il reddito familiare equivalente è dell'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE), ed è pari alla somma di diversi coefficienti individuali (1 per il primo adulto, 0,5 per ogni altro adulto e 0,3 per ogni componente di età inferiore a 14 anni).

Titolo di studio: rappresenta il grado di istruzione degli individui ed è derivato prendendo come base il titolo di studio dichiarato al Censimento della Popolazione del 2011, aggiornato con i titoli di studio conseguiti successivamente al 2011 registrati nella Base dati integrata su istruzione e titoli di studio (BIT).

Partecipazione a percorsi formativi scolastici o universitari: è determinata dalla presenza dei componenti della famiglia anagrafica nell'Anagrafe nazionale degli studenti delle scuole nell'anno scolastico $t/t+1$ o nell'Anagrafe nazionale degli studenti universitari - Iscritti e iscrizioni nell'anno accademico $t/t+1$. Non tutti i percorsi sono tracciabili nei dati amministrativi.

1.3 Misure di qualità dell'output rispetto alle statistiche ufficiali e indicatori di qualità

Questo paragrafo presenta alcune misure di qualità dell'output riferite alla base dati. In particolare, vengono presentati i principali risultati emersi dai confronti effettuati tra la base dati e le fonti ufficiali.

Per quanto riguarda la tempestività, la data di pubblicazione dei dati è $t+2$ anni rispetto alla data di riferimento. Sul divario temporale incide particolarmente l'acquisizione di alcune fonti, soprattutto quelle di natura fiscale.

Relativamente alle tempistiche di produzione della base dati, si segnala che il tempo complessivo (comprensivo dei controlli) per realizzare la fase di integrazione è di circa quattordici giorni lavorativi nell'ipotesi di condizioni ottimali, ovvero nel caso in cui non vengano riscontrati errori e non si rendano necessarie modifiche sui dati. Tale tempo aumenta in funzione della possibilità di rieseguire la procedura al momento in cui viene richiesta, nel caso in cui la pianificazione delle attività effettuata venga modificata.

Per valutare la qualità delle informazioni disponibili nella base prima dell'armonizzazione nel SIR, sono stati effettuati dei confronti rispetto alle statistiche ufficiali. Il prospetto 1.2 riporta l'elenco degli aggregati sottoposti ad analisi di coerenza. Nei sotto-paragrafi che seguono vengono illustrate le principali evidenze emerse.

Prospetto 1.2 - Aggregati oggetto di analisi di coerenza rispetto a fonti ufficiali di riferimento

FONTI DI RIFERIMENTO	ANALISI DI COERENZA EFFETTUATE
Istat, Rilevazione annuale del movimento e calcolo della popolazione residente e della popolazione residente straniera	Popolazione residente per sesso e classe di età; Popolazione residente per comune; Popolazione residente straniera per comune.
Istat, Rilevazione sulle forze di lavoro	Popolazione per titolo di studio; indicatori regionali su % di laureati e % di Neet.
Istat, Multiscopo sulle famiglie: aspetti della vita quotidiana	Numero di famiglie; famiglie per numero di componenti; tipologia familiare.
Mef, Analisi statistiche dei dati derivanti dalle dichiarazioni dei redditi delle persone fisiche	Percettori di reddito; ammontare del reddito per regione.
Istat, Indagine su reddito e condizioni di vita (Eu-Silc)	Reddito medio; indicatori regionali su «rischio povertà» e bassa intensità lavorativa familiare.

Fonte: Elaborazione degli autori



1.3.1 Coerenza rispetto alla popolazione residente

I dati ufficiali utilizzati per il confronto con la Popolazione residente (e con la Popolazione residente straniera) provengono dalla Rilevazione annuale del movimento e calcolo della popolazione residente e della popolazione residente straniera, 1 gennaio 2018 - Istat. I dati si riferiscono a tutti i comuni italiani. La rilevazione include nella popolazione residente anche gli individui residenti in convivenza, mentre la popolazione di riferimento di “Condizioni socio-economiche delle famiglie - Archimede” contiene solo i residenti in famiglia.

Il totale della popolazione residente al 1 gennaio 2018 secondo il riferimento ammonta a 60.483.973 individui (Tavola 1.1). La base (solo residenti in famiglia) ne conta 60.221.330 (262 mila unità in meno, pari a una differenza dello 0,4 per cento rispetto al bilancio demografico). Nelle varie classi di età la differenza percentuale diventa trascurabile. L’attesa sotto-copertura che si osserva rispetto a questa fonte è in buona parte dovuta alla diversa popolazione di riferimento delle due fonti costituita dalle persone in convivenza.

La coerenza rispetto alle variabili Sesso e Classe di età risulta elevata. In media, l’errore medio assoluto percentuale (MAPE⁴) della base rispetto alla Rilevazione Istat è dello 0,55 per cento.

Tavola 1.1 - Popolazione residente per sesso e classe di età, Rilevazione Istat e “Condizioni socio-economiche delle famiglie - Archimede”. Anno 2017 (valori assoluti)

CLASSE DI ETÀ	Maschi	Femmine	Missing	Totale
RILEVAZIONE ISTAT				
0-14	4.157.602	3.922.574	0	8.080.176
15-34	6.435.624	6.092.600	0	12.528.224
35-64	12.918.543	13.312.667	0	26.231.210
65 e più	5.915.838	7.728.525	0	13.644.363
Totale	29.427.607	31.056.366	0	60.483.973
CONDIZIONI SOCIO-ECONOMICHE DELLE FAMIGLIE - ARCHIMEDE				
0-14	4.159.783	3.924.466	0	8.084.249
15-34	6.356.967	6.094.480	1	12.451.448
35-64	12.886.228	13.300.352	0	26.186.580
65 e più	5.866.430	7.582.820	0	13.449.250
Totale	29.269.408	30.902.118	1	60.171.527

Fonte: Elaborazione degli autori

Dal confronto con la popolazione a livello comunale si osserva che l’errore medio della base rispetto ai dati ufficiali è pari allo 0,8 per cento. Relativamente ai soli cittadini stranieri, l’errore medio sale al 9,4 per cento. Questo dato indica una qualità inferiore dell’informazione sulla Cittadinanza in “Condizioni socio-economiche delle famiglie - Archimede”: tale variabile è per ora acquisita dalle Liste anagrafiche comunali e contiene più di mille *missing* oltre ad errori di codifica.

1.3.2 Coerenza rispetto al grado di istruzione

La Rilevazione sulle forze di lavoro (Rfl) rappresenta la principale fonte di informazione statistica sul mercato del lavoro italiano. Le tavole seguenti riportano i confronti effettuati relativamente al grado di istruzione rilevato dall’indagine.

4 *Mean absolute percentage error*: è calcolato come media delle differenze in valore assoluto tra i valori degli indicatori calcolati sulle due fonti.

1. Base integrata “Condizioni socio economiche delle famiglie”

La distribuzione della popolazione di 15 anni e più per titolo di studio in “Condizioni socio-economiche delle famiglie - Archimede” evidenzia (nella composizione percentuale) una buona somiglianza con quella della Rfl (Tavola 1.2). Tuttavia, si registra una leggera sovra-copertura nella stima. L'errore medio della base rispetto alla Rfl è del 4 per cento.

Tavola 1.2 - Popolazione 15 anni e oltre per titolo di studio, Rilevazione sulle forze di lavoro (media annua) e “Condizioni socio-economiche delle famiglie - Archimede”. Anno 2017 (valori assoluti, dati in migliaia)

PAESE DI CITTADINANZA	Rilevazione sulle forze di lavoro		Condizioni socio-economiche delle famiglie - Archimede		Differenza %
	Valori assoluti	Valori %	Valori assoluti	Valori %	
Licenza di scuola elementare, nessun titolo di studio	9.238	17,7	9.187	18,5	-0,6
Licenza di scuola media	16.804	32,3	16.045	32,2	-4,5
Diploma 2-3 anni (qualifica professionale)	2.877	5,5	2.745	5,5	-4,6
Diploma 4-5 anni (maturità)	15.779	30,3	14.668	29,5	-7,0
Laurea e post-laurea	7.355	14,1	7.132	14,3	-3,0
Totale	52.053	100,0	49.778	100,0	-4,4

Fonte: Elaborazione degli autori

L'indicatore sulla percentuale di laureati di età compresa tra 30 e 34 anni, a livello di regione, calcolato su entrambe le fonti mostra una buona coerenza nella distribuzione territoriale, anche se la base sovra-stima leggermente il dato a livello Italia e nella maggior parte delle regioni (Tavola 1.3). L'errore medio di “Condizioni socio-economiche delle famiglie - Archimede” rispetto alla Rfl è pari a 3,5.

Tavola 1.3 - Popolazione di età compresa tra 30 e 34 anni con istruzione terziaria per regione, Rilevazione sulle forze di lavoro e “Condizioni socio-economiche delle famiglie - Archimede”. Anno 2017 (incidenza percentuale sul totale della popolazione di età compresa tra 30 e 34 anni)

REGIONE	Rilevazione sulle forze di lavoro	Condizioni socio-economiche delle famiglie - Archimede
Piemonte	26,4	29,1
Valle d'Aosta/Vallée d'Aoste	25,2	29,5
Liguria	23,7	30,7
Lombardia	33,7	32,4
Bolzano/Bozen	24,6	22,6
Trento	33,6	34,6
Veneto	27,6	29,9
Friuli-Venezia Giulia	28,7	31,5
Emilia-Romagna	29,9	32,7
Toscana	28,3	30,6
Umbria	29,7	32,1
Marche	33,0	33,5
Lazio	30,1	34,2
Abruzzo	25,8	32,9
Molise	26,1	35,9
Campania	21,4	25,1
Puglia	22,2	25,7
Basilicata	29,2	30,9
Calabria	20,7	28,9
Sicilia	19,1	23,8
Sardegna	23,6	23,8
ITALIA	26,9	29,5

Fonte: Elaborazione degli autori

1.3.3 Coerenza rispetto al reddito familiare

L'indagine sul reddito e le condizioni di vita (Eu-Silc) risponde alla richiesta di informazione statistica su argomenti come redditi, povertà, esclusione sociale, deprivazione (Donatiello, 2011). L'indagine è condotta su un campione di circa 29 mila famiglie (per un totale di quasi 70 mila individui).

Tavola 1.4 - Reddito mediano equivalente per alcune caratteristiche delle famiglie, Indagine sul reddito e le condizioni di vita (Eu-Silc) e "Condizioni socio-economiche delle famiglie - Archimede". Anno 2017

TIPO DI FAMIGLIA	Reddito mediano netto	Reddito mediano lordo
	Indagine sul reddito e le condizioni di vita Eu-Silc	Condizioni socio-economiche delle famiglie - Archimede
Tutte le famiglie	16.844	17.417
Famiglie mono-componente	16.322	16.972

Fonte: Elaborazione degli autori

La Tavola 1.4 riporta il valore del reddito mediano nelle due fonti. Nel confronto è, tuttavia, opportuno tenere presente che la base dati include soltanto voci di reddito al lordo della tassazione mentre per l'indagine Eu-Silc sono indicati, invece, i redditi netti. Ciò premesso i risultati sono conformi alle aspettative.

È stato inoltre calcolato l'indicatore a livello regionale relativo alla percentuale di persone che vivono in famiglie con reddito inferiore al 60 per cento del reddito mediano. L'errore di "Condizioni socio-economiche delle famiglie - Archimede" rispetto a Eu-Silc risulta pari al 6,8 per cento.

I valori dell'indicatore calcolato sui dati di Archimede risultano superiori a quelli di Eu-Silc (Tavola 1.5). Questo effetto può essere dovuto a una pluralità di fattori, anche concomitanti:

- utilizzo in Archimede del reddito lordo: in Eu-Silc la percentuale di famiglie con un reddito equivalente inferiore al 60 per cento del reddito mediano equivalente viene calcolata utilizzando il reddito netto e risente, quindi, dell'effetto redistributivo dell'imposta, mentre i redditi in Archimede sono al lordo della tassazione (da ciò deriva una maggiore disuguaglianza nei redditi e una povertà relativa più diffusa);
- differenze definitorie delle diverse voci di reddito tra le due fonti;
- sotto-copertura o errori di quantificazione del reddito nelle fonti amministrative.

1. Base integrata “Condizioni socio economiche delle famiglie”

Tavola 1.5 - Individui che vivono in famiglie con reddito inferiore al 60 per cento del reddito mediano per regione, Indagine sul reddito e le condizioni di vita (Eu-Silc) e “Condizioni socio-economiche delle famiglie - Archimede”. Anno 2017 (valori percentuali)

REGIONE	Indagine sul reddito e le condizioni di vita	Condizioni socio-economiche delle famiglie -
	Eu-Silc	Archimede
Piemonte	14,2	19,4
Valle d'Aosta/Vallée d'Aoste	12,0	16,4
Liguria	14,0	21,8
Lombardia	11,1	17,9
Bolzano/Bozen	9,2	12,6
Trento	15,3	14,6
Veneto	11,0	15,8
Friuli-Venezia Giulia	8,2	15,4
Emilia-Romagna	10,1	16,2
Toscana	14,4	19,6
Umbria	12,5	21,4
Marche	11,7	19,4
Lazio	19,3	28,9
Abruzzo	18,7	27,7
Molise	23,5	34,4
Campania	41,4	45,1
Puglia	26,8	37,4
Basilicata	30,1	34,3
Calabria	32,7	45,8
Sicilia	40,7	46,6
Sardegna	27,1	33,7
ITALIA	20,3	27,1

Fonte: Elaborazione degli autori

1.3.4 Coerenza rispetto al reddito individuale

Per quanto riguarda i redditi degli individui, una delle fonti di riferimento è costituita dalle analisi statistiche dei dati derivanti dalle dichiarazioni dei redditi delle persone fisiche a cura del Dipartimento delle Finanze del Ministero dell'economia e finanze (Mef).

La distribuzione territoriale del reddito risulta in linea nelle due fonti.

La leggera sovrastima del numero dei percettori e dell'ammontare di reddito nella base che si osserva rispetto a questa fonte (Tavola 1.6) può essere, almeno in parte, dovuta a una differenza nella popolazione di riferimento (il numero di percettori di reddito secondo le statistiche fiscali è riferito a tutti i contribuenti, ovvero percettori di redditi imponibili, mentre nella base è riferito ai percettori – anche di redditi non imponibili – residenti in famiglia) e delle variabili di interesse.

L'errore medio della base rispetto alle statistiche fiscali è del 2,7 per cento per quanto riguarda la frequenza dei percettori di reddito e del 5,3 per cento per quanto riguarda l'ammontare del reddito.

Tavola 1.6 - Percettori di reddito, ammontare di reddito e reddito medio per regione. Statistiche fiscali Mef - Dipartimento delle finanze e "Condizioni socio-economiche delle famiglie - Archimede". Anno 2017 (valori assoluti, ammontare e media in migliaia di euro)

REGIONE	Reddito complessivo Mef			Reddito complessivo Condizioni socio-economiche delle famiglie - Archimede		
	Frequenza	Ammontare	Media	Frequenza	Ammontare	Media
Piemonte	3.141.112	69.824.022	22,23	3.117.805	71.853.108	23,05
Valle d'Aosta/Vallée d'Aoste	96.128	2.126.759	22,12	93.785	2.187.279	23,32
Liguria	1.163.588	25.505.645	21,92	1.121.646	25.874.488	23,07
Lombardia	7.111.447	175.760.663	24,72	7.048.095	178.842.477	25,37
Trento	411.087	8.918.574	21,70	392.196	8.986.951	22,91
Bolzano/Bozen	423.272	10.093.088	23,85	388.441	10.122.950	26,06
Veneto	3.556.723	77.763.220	21,86	3.481.307	79.366.557	22,80
Friuli Venezia Giulia	920.863	20.058.516	21,78	886.411	20.279.442	22,88
Emilia Romagna	3.337.935	76.354.781	22,87	3.266.752	77.911.218	23,85
Toscana	2.694.909	57.332.158	21,27	2.665.544	59.277.628	22,24
Umbria	619.609	12.067.494	19,48	626.105	12.812.024	20,46
Marche	1.107.196	21.420.469	19,35	1.102.320	22.434.639	20,35
Lazio	3.819.697	86.612.948	22,68	3.872.127	89.929.073	23,22
Abruzzo	894.292	15.562.099	17,40	896.375	16.611.299	18,53
Molise	207.842	3.212.535	15,46	204.413	3.476.756	17,01
Campania	3.115.533	51.490.255	16,53	3.318.897	56.925.422	17,15
Puglia	2.529.867	39.535.715	15,63	2.576.624	43.241.423	16,78
Basilicata	373.031	5.755.367	15,43	375.875	6.280.644	16,71
Calabria	1.164.584	16.445.988	14,12	1.203.206	18.770.741	15,60
Sicilia	2.818.281	44.365.378	15,74	2.942.546	49.297.358	16,75
Sardegna	1.042.461	18.002.717	17,27	1.094.862	19.521.736	17,83
Non indicata	5.513	17.651	3,20	0	0	-
ITALIA	40.554.970	838.226.041	20,67	40.675.332	874.003.213	21,49

Fonte: Elaborazione degli autori

1.3.5 Coerenza rispetto alla distribuzione delle famiglie

L'Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana (AVQ), realizzata a cadenza annuale, è condotta su un campione di circa 20 mila famiglie e 50 mila individui. La Tavola 1.7 mostra i risultati dei confronti effettuati relativamente alla consistenza delle famiglie (per AVQ il periodo di riferimento dei dati è la media 2016-2017).

Tavola 1.7 - Numero di famiglie. Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana e "Condizioni socio-economiche delle famiglie - Archimede". Anno 2017 (valori assoluti in migliaia)

	Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana	Condizioni socio-economiche delle famiglie - Archimede
Italia (v.a in migliaia)	25.494	26.003
Differenza assoluta		509
Differenza percentuale		2

Fonte: Elaborazione degli autori

L'errore medio della base rispetto all'indagine relativamente alla composizione percentuale per numero di componenti delle famiglie è dello 0,05 per cento a fronte di una sovrastima complessiva del 2 per cento del numero di famiglie. In particolare, la differenza maggiore si riscontra tra le famiglie con un solo componente: la base presenta una incidenza maggiore di queste famiglie (34,8 per cento del totale) rispetto all'indagine (31,9 per cento) (Tavola 1.8).

1. Base integrata “Condizioni socio economiche delle famiglie”

Tavola 1.8 - Famiglie per numero di componenti. Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana e “Condizioni socio-economiche delle famiglie - Archimede”. Anno 2017 (composizione percentuale)

NUMERO DI COMPONENTI	Indagine Multiscopo sulle famiglie: aspetti della vita quotidiana	Condizioni socio-economiche delle famiglie - Archimede	Differenza
Un componente	31,9	34,8	2,9
Due componenti	27,5	26,6	-0,9
Tre componenti	19,6	18,6	-1,0
Quattro componenti	15,7	14,7	-1,0
Cinque o più componenti	5,3	5,3	0,0
Totale	100,0	100,0	

Fonte: Elaborazione degli autori

Ciò è presumibilmente dovuto alla qualità del dato anagrafico dal quale queste informazioni sono tratte. Infatti, la tipologia e la numerosità delle famiglie sono strettamente dipendenti dalla qualità delle informazioni delle liste anagrafiche da cui vengono derivate e possono essere inficiate da mancate regolarizzazioni anagrafiche di situazioni familiari mutate o da comportamenti opportunistici degli individui (ad esempio le residenze fittizie che non consentono di definire in modo corretto le unità familiari).

1.4 Analisi di corrispondenza tra informazioni del Registro base degli Individui (RBI) e quelle contenute nella base

Nell’ottica di armonizzare i contenuti informativi della base dati nel Sistema integrato dei registri, si è ritenuto opportuno effettuare una valutazione della coerenza dei contenuti informativi della base dati “Condizioni socio-economiche delle famiglie - Archimede” con gli analoghi contenuti derivanti dal Registro statistico di base degli individui, delle famiglie e delle convivenze. Tale analisi, condotta per l’anno di riferimento 2017, si configura come una prima valutazione dell’impatto della sostituzione di alcuni contenuti informativi.

La ricognizione preliminare ha previsto lo svolgimento delle seguenti attività:

- *mapping* dei contenuti informativi nelle due fonti, attraverso il quale si sono tratte le prime indicazioni sulla sostituibilità della popolazione di input e di alcune variabili;
- analisi di coerenza delle classificazioni delle variabili delle due fonti;
- controllo di eventuali disallineamenti temporali nel periodo di riferimento delle variabili.

Il prospetto 1.3 riporta l’esito del *mapping* delle principali variabili di “Condizioni socio-economiche delle famiglie - Archimede” in RBI (riferito all’annualità 2017).

Prospetto 1.3 - Mapping delle principali variabili di “Condizioni socio-economiche delle famiglie - Archimede” nel Registro base degli individui

TIPO DI VARIABILE	DESCRIZIONE VARIABILE	Fonte della variabile in condizioni socio-economiche delle famiglie - Archimede E	ETICHETTA DELLA VARIABILE IN RBI
Variabili individuali anagrafiche	Sesso	Liste anagrafiche comunali	Sesso
	Data di nascita	Liste anagrafiche comunali	Data_nascita
	Cittadinanza	Liste anagrafiche comunali	Cod_cittadinanza
	Stato civile	Liste anagrafiche comunali	Staciv
	Relazione di parentela	Liste anagrafiche comunali	Rel_par
	Codice Istat stato estero di nascita	Liste anagrafiche comunali	Cod_paese_nascita
Variabili familiari	Codice famiglia	Liste anagrafiche comunali	Cod_famiglia
	Numero componenti della famiglia	Liste anagrafiche comunali	(Ricalcolabile sulla base di cod_famiglia)
Variabili individuali di residenza anagrafica	Codice provincia di residenza	Liste anagrafiche comunali	Cod_pro_residenza
	Codice comune di residenza	Liste anagrafiche comunali	Cod_com_residenza
Variabili relative al titolo di studio	Grado di istruzione	Censimento popolazione 2011 + BIT	Cod_grado_istruzione1

Fonte: Elaborazione degli autori

Successivamente sono stati eseguiti dei controlli su dati aggregati, relativamente a:

- numerosità della popolazione di riferimento (individui e famiglie), fino al livello territoriale comunale, relativamente alla popolazione residente in famiglia;
- distribuzione secondo variabili demografiche individuali (in particolare Data di nascita, Sesso, Cittadinanza) con disaggregazione territoriale.

Infine, è stata svolta un'analisi a livello di micro-dati delle eventuali incoerenze tra registro da un lato e base dati “Condizioni socio-economiche delle famiglie - Archimede” dall'altro. Nel dettaglio:

- analisi della popolazione in comune alle due fonti;
- analisi delle coerenze/incoerenze tra le fonti, per gli individui in comune, sulle seguenti variabili: sesso, data di nascita, comune di residenza, cittadinanza, grado di istruzione.

1.4.1 Controlli su macro-dati

Al fine di effettuare i confronti sulla stessa popolazione di riferimento, su RBI sono stati selezionati solamente gli individui in vita e residenti in famiglia. La Tavola 1.9 riporta il numero complessivo di comuni, individui e famiglie presenti nelle due fonti.

Tavola 1.9 - Numero di Comuni, di individui residenti in famiglia e di famiglie. “Condizioni socio economiche delle famiglie - Archimede” e Registro base degli individui. Anno 2017

UNITÀ	Condizioni socio-economiche delle famiglie - Archimede	Differenza % con RBI
Comuni	7.954	-0,06
Individui	60.171.527	0,09
Famiglie	26.002.906	0,03

Fonte: Elaborazione degli autori

La base risulta leggermente sovra-coperta rispetto a RBI per il numero di individui (+0,09 per cento) e di famiglie (+0,03 per cento).

La Tavola 1.10 riporta, per alcune distribuzioni comunali, la media delle differenze relative percentuali, ovvero l'errore medio con cui i valori della base dati si discostano dai valori di RBI, che rappresentano quelli di riferimento. In generale, si osserva un livello molto basso di discrepanza tra le due fonti.

1. Base integrata “Condizioni socio economiche delle famiglie”

Tavola 1.10 - Mean absolute percentage error tra le distribuzioni a livello comunale di unità di interesse di “Condizioni socio-economiche delle famiglie - Archimede” e Registro base degli individui. Anno 2017

UNITÀ DI INTERESSE/TERRITORIO	Mean absolute percentage error (MAPE)
Individui residenti in famiglia per comune	0,16
Individui stranieri residenti in famiglia per comune	1,35
Individui italiani residenti in famiglia per comune	0,11
Famiglie anagrafiche per comune	5,37

Fonte: Elaborazione degli autori

1.4.2 Controlli su micro-dati

La Tavola 1.11 riporta le numerosità degli individui presenti in entrambe le fonti e degli individui presenti in una fonte ma assenti nell'altra. Si osserva che la sovrapposizione degli individui residenti in famiglia e in vita al 31 dicembre è totale. Gli individui in comune sono, infatti, 60 milioni e 44 mila, pari più del 99 per cento degli individui in vita residenti in famiglia del RBI.

Tavola 1.11 - Individui in comune alle due fonti, individui presenti in una sola fonte. Anno 2017 (valori assoluti e incidenza percentuale)

Individui in comune alle due fonti	60.044.274
<i>Incidenza percentuale su tot. individui residenti in famiglia in RBI</i>	99,9
<i>Incidenza percentuale su tot. individui in Condizioni socio-economiche delle famiglie – Archimede</i>	99,8
Individui presenti solo in RBI (residenti in famiglia) <i>Incidenza percentuale su tot. individui residenti in famiglia in RBI</i>	0,1
Individui presenti solo in Condizioni socio-economiche delle famiglie - Archimede <i>Incidenza percentuale su tot. individui nella base</i>	0,2
Totale individui Condizioni socio-economiche delle famiglie - Archimede	60.171.527

Fonte: Elaborazione degli autori

Sul sotto-insieme degli individui in comune alle due fonti è stata analizzata la concordanza rispetto alle seguenti variabili: Data di nascita, Sesso, Cittadinanza, Grado di istruzione, Comune di residenza.

La variabile Data di nascita mostra un grado di concordanza pari al 99,99 per cento del totale degli individui in comune. Per circa 5 mila individui si osserva invece una discordanza (Tavola 1.12).

Tavola 1.12 - Individui per concordanza della variabile Data di nascita . Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DELLA DATA DI NASCITA	Valori assoluti	Composizione %
Stessa data di nascita	60.038.999	99,99
Diversa data di nascita	5.275	0,01
Totale individui in comune	60.044.274	100,0

Fonte: Elaborazione degli autori

Anche per quanto riguarda il sesso degli individui si osserva un livello di coerenza pressoché totale tra il registro e la base, con soli 133 individui con modalità discordante (Tavola 1.13).

Tavola 1.13 - Individui per concordanza della variabile Sesso. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DEL SESSO	Valori assoluti	Composizione %
Stesso sesso	60.044.140	100
Sesso diverso	133	0
Sesso mancante su Condizioni socio-economiche delle famiglie - Archimede	1	0
Totale individui in comune	60.044.274	100

Fonte: Elaborazione degli autori

Rispetto alle altre variabili considerate, la cittadinanza sembra essere quella per cui la sostituzione dell'informazione di "Condizioni socio-economiche delle famiglie - Archimede" con quella di RBI apporterebbe un incremento di qualità più consistente. Analizzando nel dettaglio i singoli codici di cittadinanza (secondo la classificazione degli stati esteri Istat in entrambe le fonti), si osserva, infatti, una discordanza per 17.306 individui. Questa discrepanza è da attribuire al mancato trattamento della variabile cittadinanza sulle Liste anagrafiche comunali (la fonte attraverso la quale viene valorizzata questa informazione sulla base) che quindi presenta codici errati o non più validi.

Tavola 1.14 - Individui per concordanza della variabile Cittadinanza. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DELLA CITTADINANZA	Valori assoluti	Composizione %
Stesso codice cittadinanza	60.025.846	99,97
Codice cittadinanza diverso	17.306	0,03
Cittadinanza mancante in almeno una fonte	1.122	0,0
Totale individui in comune	60.044.274	100,0

Fonte: Elaborazione degli autori

Anche la concordanza sulla variabile relativa al comune di residenza tra le due fonti è molto alta, con soli 2.442 individui per cui si osserva una discordanza (Tavola 1.15).

Tavola 1.15 - Individui per concordanza della variabile Comune di residenza. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DEL COMUNE DI RESIDENZA	Valori assoluti	Composizione %
Stesso comune di residenza	60.041.832	100,0
Comune di residenza diverso	2.442	0,0
Totale individui in comune	60.044.274	100,0

Fonte: Elaborazione degli autori

Il confronto tra il grado di istruzione registrato nelle due fonti sconta il diverso riferimento temporale di una delle fonti di input utilizzate (la Base dati integrata su Istruzione e Titoli di studio - BIT). Infatti, per la versione 2017 del registro è stata utilizzata la BIT riferita al 2016, mentre per "Condizioni socio-economiche delle famiglie - Archimede" la BIT 2017. Questo comporta uno slittamento verso il grado di istruzione superiore per una quota di popolazione nella base Archimede. Al netto di questa differenza, si registra un buon grado di concordanza. Si hanno oltre 46,5 milioni di individui con lo stesso grado di istruzione sulle due fonti (Tavola 1.16). Considerando come coerenti anche i casi di "slittamento" (ovvero gli individui con un grado di istruzione immediatamente superiore in Condizioni socio-economiche delle famiglie rispetto a quello del registro), la coerenza sfiora il 90 per cento. Si hanno, inoltre, oltre 5 milioni di individui per cui il grado di istruzione è mancante o non dovuto in entrambe le fonti. Si tratta di bambini di età inferiore ai 6 anni o di individui non presenti in BIT e non censiti nel 2011.

1. Base integrata “Condizioni socio economiche delle famiglie”

Tavola 1.16 - Individui per concordanza della variabile Grado di istruzione. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DEL GRADO DI ISTRUZIONE	Valori assoluti	Composizione %
Stesso grado di istruzione	46.543.211	77,5
Slittamenti dovuti a diverso periodo di riferimento	7.389.170	12,3
Missing in entrambe le fonti o non dovuti	5.253.715	8,7
Incoerenze	858.178	1,4
Totale	60.044.274	100,0

Fonte: Elaborazione degli autori

1.5 Riflessioni conclusive, punti aperti e criticità

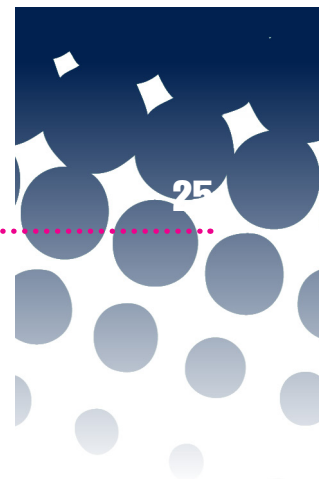
I risultati hanno mostrato che la base “Condizioni socio-economiche delle famiglie - Archimede” ha già in partenza un livello di coerenza informativa elevato con RBI, tale che i miglioramenti di qualità dell’output saranno generalmente contenuti con la sostituzione dell’informazione della base con quella da registro. In particolare, l’errore medio (MAPE) con cui i valori della base si discostano dai valori di RBI è molto basso per tutte le distribuzioni comunali confrontate: per gli individui residenti in famiglia l’errore medio che si andrà a recuperare è dello 0,16 per cento. Anche i confronti sui micro-dati sulle variabili demografiche in comune alle due fonti mostrano livelli di concordanza superiori al 99 per cento. Solo per la cittadinanza la sostituzione dell’informazione della base con quella di RBI apporterebbe un incremento di qualità leggermente più consistente.

L’analisi del processo effettuata ha messo in luce che l’armonizzazione nel SIR è però funzionale a miglioramenti in termini di semplificazione e aumenti di efficienza, di minor numero di fonti di input da utilizzare e di razionalizzazione delle fasi di lavorazione dei dati.

Tuttavia, al momento, data la sostituibilità incompleta di contenuti informativi della base con analoghi contenuti da SIR (si pensi ad esempio a registri tuttora in fase di implementazione, come quello sui redditi), non si realizzerebbe nel breve periodo un incremento nella tempestività del rilascio dei dati (t+2). Infatti, per alcuni input fondamentali per la realizzazione della base l’alimentazione da SIM non può essere sostituita con l’alimentazione da SIR.

Inoltre, è necessario considerare che il sistema dei registri si configura come un sistema in evoluzione continua. In alcuni casi è possibile prevedere una modellazione “aperta” ad altri registri, in altri casi si sottolinea che l’introduzione di modifiche nel corso del tempo sui contenuti informativi delle principali variabili da registro usate per la sostituzione (es. alcune variabili di RBI, come il titolo di studio) in termini di classificazioni, riferimenti temporali, ecc. ha un impatto anche sulla base. La messa a disposizione da parte dell’Istituto di altri registri, utilizzabili come input per la base, comporterà anche una rivalutazione dell’eventuale incremento nella tempestività del rilascio dei dati.

Si segnala, infine, che la definizione dell’*as-is* risulta suscettibile di piccole modifiche, indipendentemente dall’integrazione con i registri in quanto, occasionalmente, vengono apportati cambiamenti al processo di costruzione della base, in relazione a variazioni dei contenuti informativi di alcuni archivi o alla loro possibile dismissione (ad esempio a seguito di modifiche normative).



2. BASE INTEGRATA “POPOLAZIONE INSISTENTE GIORNALIERA PER STUDIO E LAVORO”¹

2.1 Premessa

Nel capitolo verranno illustrati i principali risultati dello studio di fattibilità sulla produzione armonizzata tra la base integrata di microdati amministrativi “Popolazione insistente giornaliera per studio e lavoro - Archimede” ed il Sistema Integrato dei Registri (SIR). Nella prima parte vengono descritte tutte le attività che hanno condotto alla redazione dello stato attuale del processo di produzione (Paragrafo 2.2) e si presentano gli schemi rappresentanti le procedure in essere (Paragrafo 2.3). Successivamente, si illustra l'analisi della corrispondenza tra le informazioni e i concetti della base “Popolazione insistente giornaliera per studio e lavoro - Archimede” e quelle del Registro base degli individui, delle famiglie e delle convivenze (RBI), con l'individuazione delle variabili mutuabili e/o sostituibili dal Registro (Paragrafo 2.4). Si presentano poi alcuni risultati relativi a misure di qualità della base rispetto alle statistiche ufficiali (Paragrafo 2.5) e i principali risultati del confronto tra le variabili comuni alla base e RBI (Paragrafo 2.6). Alcune conclusioni e una prospettiva sugli sviluppi futuri concludono il capitolo (Paragrafo 2.7).

2.2 Informazioni per la redazione dell'analisi *as-is* del processo

La prima fase delle attività ha previsto l'esame delle caratteristiche tecniche e di architettura del processo di produzione della base² e la conseguente illustrazione dei principali elementi necessari alla modellazione del processo.

Nel dettaglio sono state condivise all'interno del gruppo di lavoro le seguenti informazioni:

- a. Le definizioni dei principali concetti e contenuti della base (definizione di Popolazione insistente diurna per lavoro, studio e residenza, Comune di origine e destinazione, Lavoratore, Studente, Studente universitario) e della popolazione di riferimento;
 - b. Le fonti acquisite e le strutture Istat interessate al rilascio delle fonti;
 - c. Le variabili acquisite dalle fonti con cenni sulla loro sostituibilità con le variabili nei registri;
 - d. Le macro fasi del processo di produzione della base: acquisizione, integrazione, trattamento dei dati;
 - e. La definizione delle tempistiche di produzione della base dati (valutazione delle tempistiche di disponibilità delle fonti di input) e della tempestività.
- Sono state, infine, fornite indicazioni su:
- f. Interconnessioni in essere e/o potenziali tra la base “Popolazione insistente giornaliera per studio e lavoro - Archimede” e le basi “Condizioni socio-economiche delle famiglie - Archimede” (Capitolo 1) e “Precarietà lavorativa - Archimede” (Capitolo 3).

¹ I Paragrafi 2.1, 2.2, 2.5 e 2.7 sono a cura di Gaia Rocchetti; il Paragrafo 2.3 è a cura di Paolo Francescangeli; i Paragrafi 2.4 e 2.6 sono a cura di Stefania Di Domenico.

² Il periodo di riferimento dei dati è il 2017.

2.3 Analisi *as-is* del processo

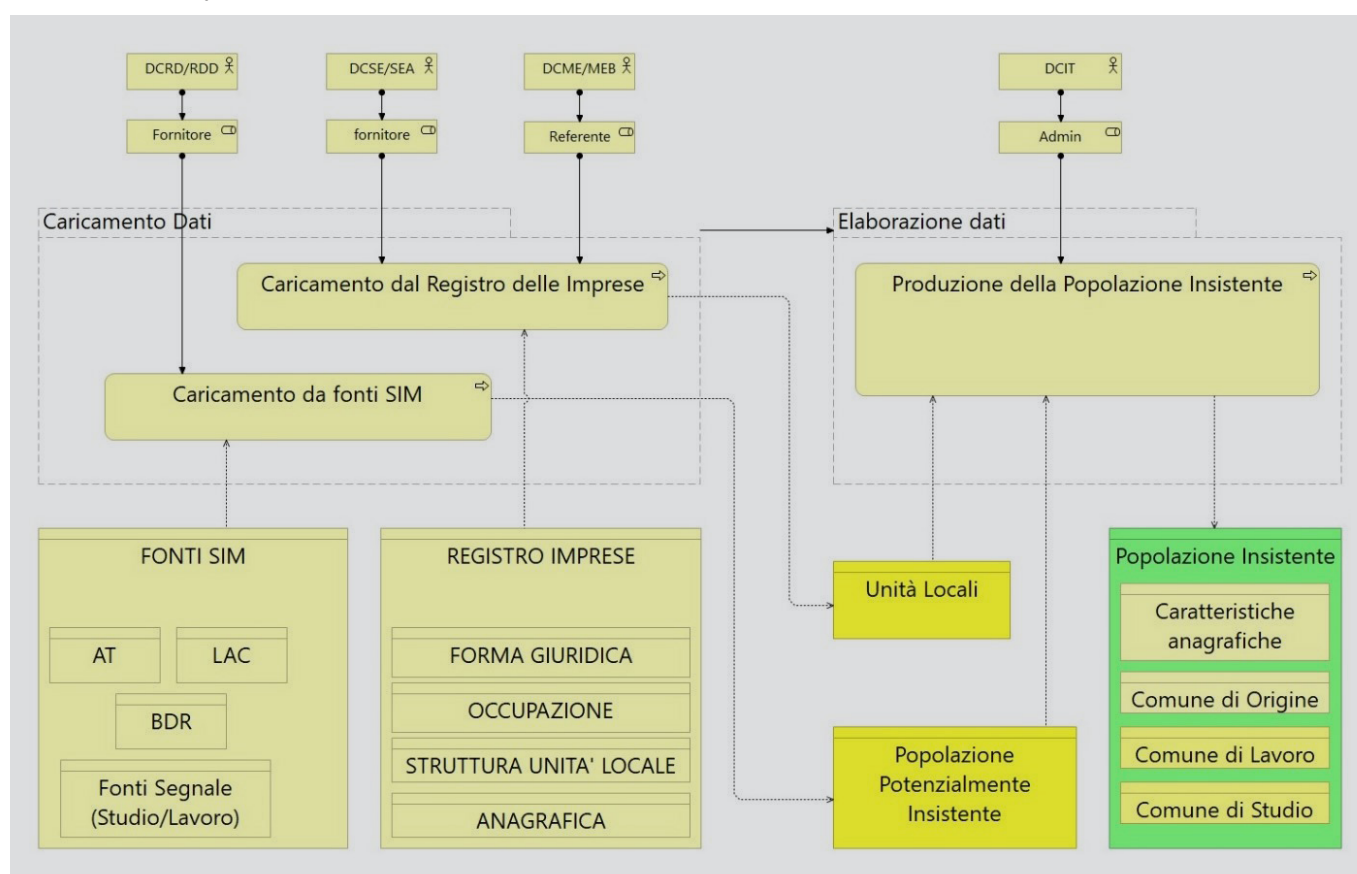
Sulle base delle informazioni condivise sono stati realizzati lo schema generale del processo e quelli di dettaglio, utilizzando il linguaggio di *Enterprise Architecture* Archimate. La modellazione così presentata ha consentito di visualizzare gli attori coinvolti, le fasi del processo, gli oggetti e le relazioni.

2.3.1 Schema generale

La Figura 2.1 riporta lo schema generale del processo. Nella parte superiore della figura sono riportati gli attori responsabili del rilascio delle fonti utilizzate in input per la costruzione della base. Tali fonti sono acquisite da:

- Direzione centrale per la raccolta dati (DCRD) che effettua una elaborazione *ad hoc* delle fonti amministrative selezionate per il processo di costruzione della base e presenti nel SIM, realizzando una base informativa integrata preliminare, e mette a disposizione una serie di viste estratte dalle fonti precedenti e da altre fonti funzionali a fasi di processo successive (es. Miur - Anagrafe nazionale degli studenti delle scuole, Miur - Anagrafe delle scuole statali e non statali, degli istituti comprensivi di ogni ordine e grado, Inail - Posizioni assicurative territoriali, Inps - Lavoratori agricoli autonomi, ecc.);

Figura 2.1 - *As-is* schema generale del processo di produzione relativo alla base dati “Popolazione insistente giornaliera per studio e lavoro - Archimede”



2. Base integrata “Popolazione insistente giornaliera per studio e lavoro”

- Direzione centrale per le statistiche economiche (DCSE) che fornisce alcune viste dall'Archivio Statistico delle imprese attive (Asia Imprese), dal Registro statistico delle unità locali (Asia Unità locali) e dal Registro statistico dell'occupazione delle imprese (Asia Occupazione).

In questa fase la Direzione centrale per la metodologia e il disegno dei processi statistici (DCME) si occupa dei primi controlli sui dati acquisiti e di alcune elaborazioni preliminari sui dati Asia.

Le fonti selezionate sono poi utilizzate dal processo Caricamento dati – nella figura, sotto i relativi attori – che si distingue nei due processi distinti Caricamento da fonti del SIM e Caricamento da Asia. In basso, sono invece rappresentate schematicamente le fonti acquisite dai due sotto-processi.

Il processo Caricamento dati fornisce in output i due prodotti intermedi: Unità locali e Popolazione potenzialmente insistente:

- Unità locali fornisce, per ogni individuo presente in Asia Unità locali, alcune informazioni sulla prestazione lavorativa dell'individuo, nonché il codice comune dell'unità locale in cui l'individuo lavora;
- Popolazione potenzialmente insistente fornisce l'insieme degli individui residenti in Italia al 31 dicembre dell'anno di riferimento e di quelli che sono presenti in una lista di fonti che li definiscono quali potenziali studenti o lavoratori sul territorio (ulteriori dettagli nel paragrafo successivo).

Gli output intermedi di Caricamento dati alimentano il processo Produzione della Popolazione insistente gestito dalla Direzione centrale per le tecnologie informatiche (DCIT) (a destra della figura). Gli output di tale processo sono: le caratteristiche anagrafiche dell'individuo, il comune di origine e il comune di destinazione (studio/lavoro).

2.3.2 Schema di dettaglio dei sotto-processi (fasi)

Ognuno dei macro-processi descritti in Figura 2.1 è stato rappresentato in modo dettagliato, descrivendo i sotto-processi (o fasi) che lo compongono. Per ogni sotto-processo sono state quindi indicate le fonti in input, le operazioni di integrazione e gli output.

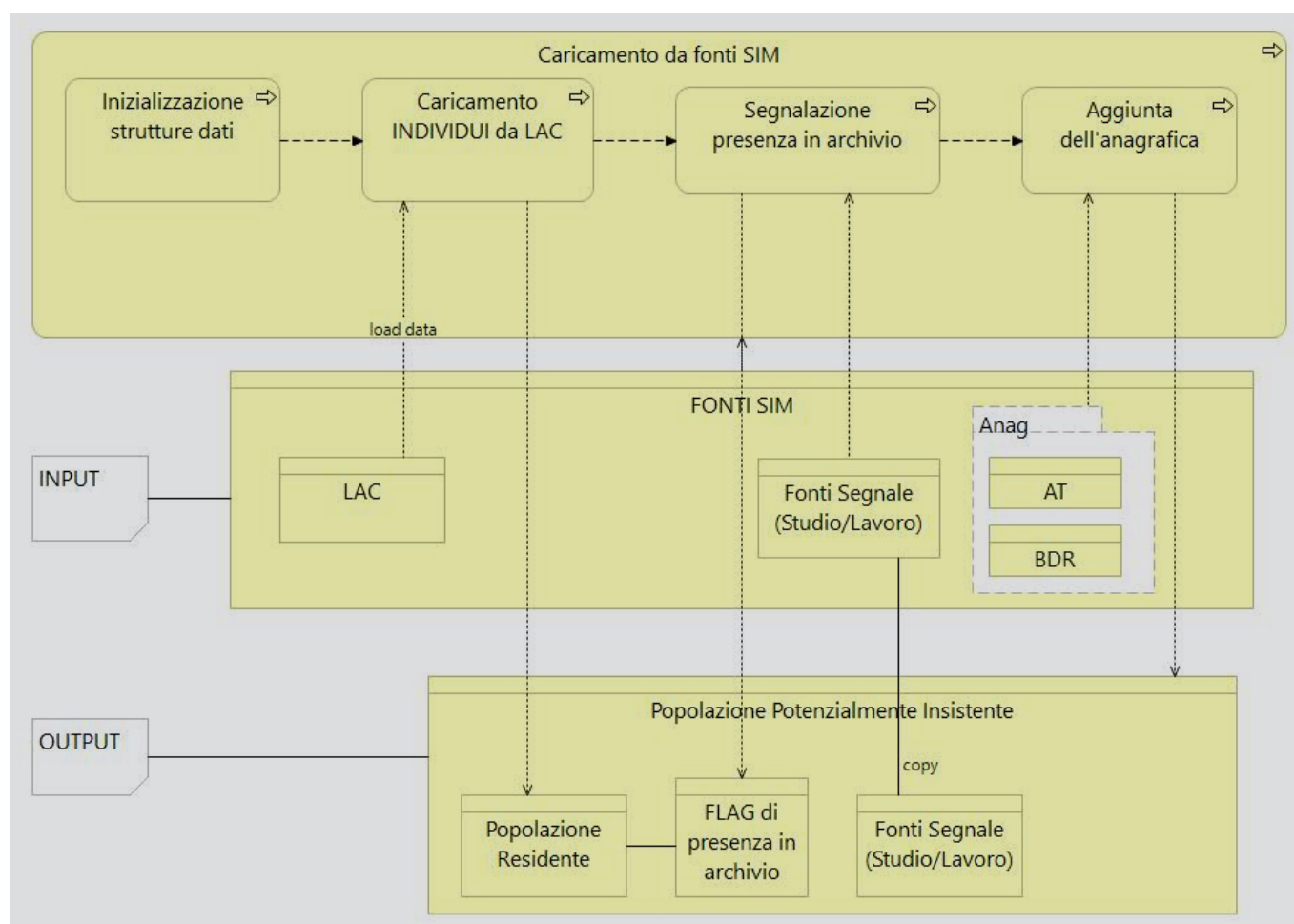
La Figura 2.2 rappresenta più nel dettaglio il processo *Caricamento fonti da Sistema integrato dei microdati*. Nella parte superiore della figura sono rappresentate le quattro fasi di cui si compone:

- inizializzazione della struttura dati;
- caricamento individui da Liste anagrafiche comunali (Lac) e contestuale caricamento del comune di residenza associato;
- segnalazione della presenza nelle fonti studio/lavoro;
- aggiunta delle informazioni anagrafiche, ovvero il Sesso, la Data e il Luogo di nascita dall'archivio Agenzia delle Entrate - Base dati anagrafica dell'anagrafe tributaria - Persone fisiche (At);

Ad ognuna di esse sono collegate le fonti in input (nella parte centrale della figura) e le variabili di output (in basso).

Le Figure 2.3 e 2.4 rappresentano, rispettivamente, lo schema del processo Caricamento da Registro delle Imprese attive e da Unità locali e Produzione della Popolazione insistente.

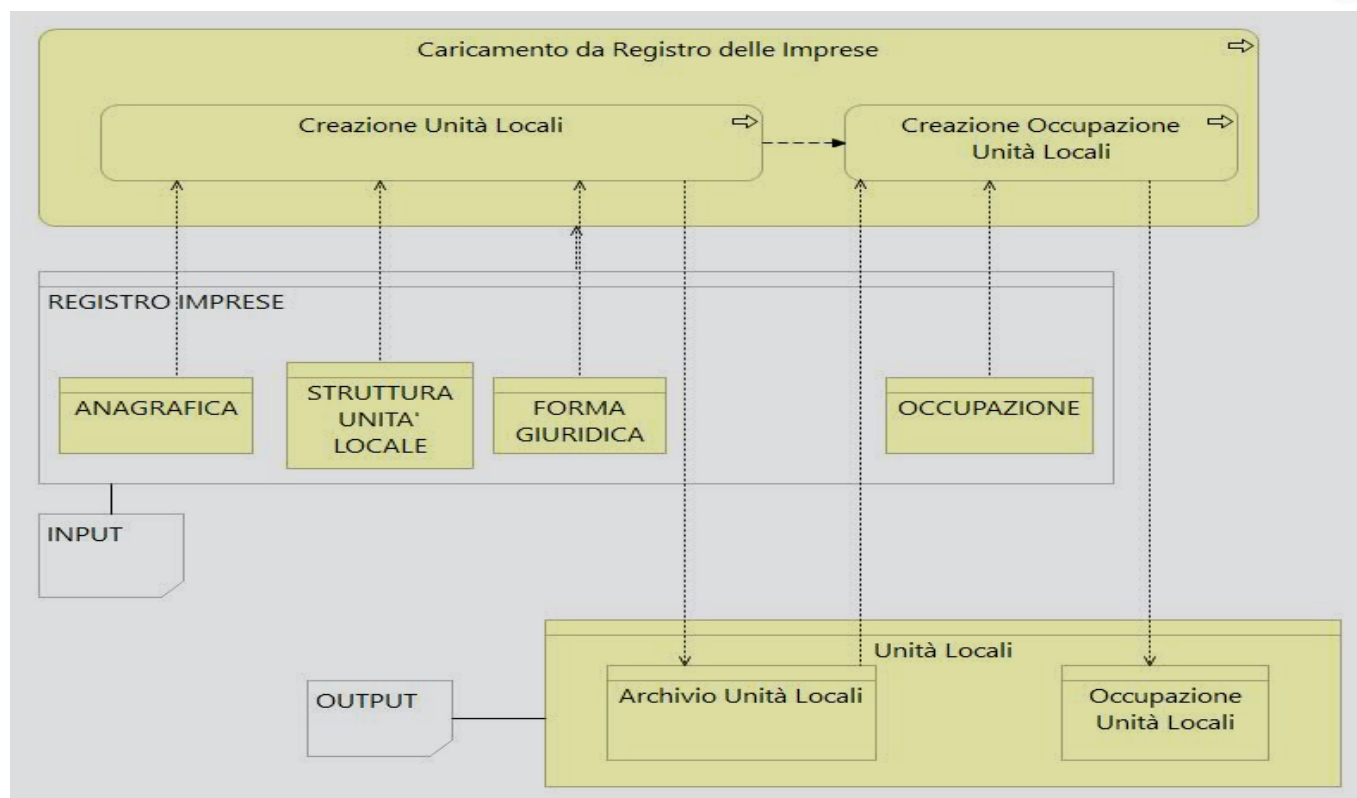
Figura 2.2 - Schema del processo Caricamento da fonti del Sistema integrato dei microdati (SIM)



Fonte: Elaborazione degli autori

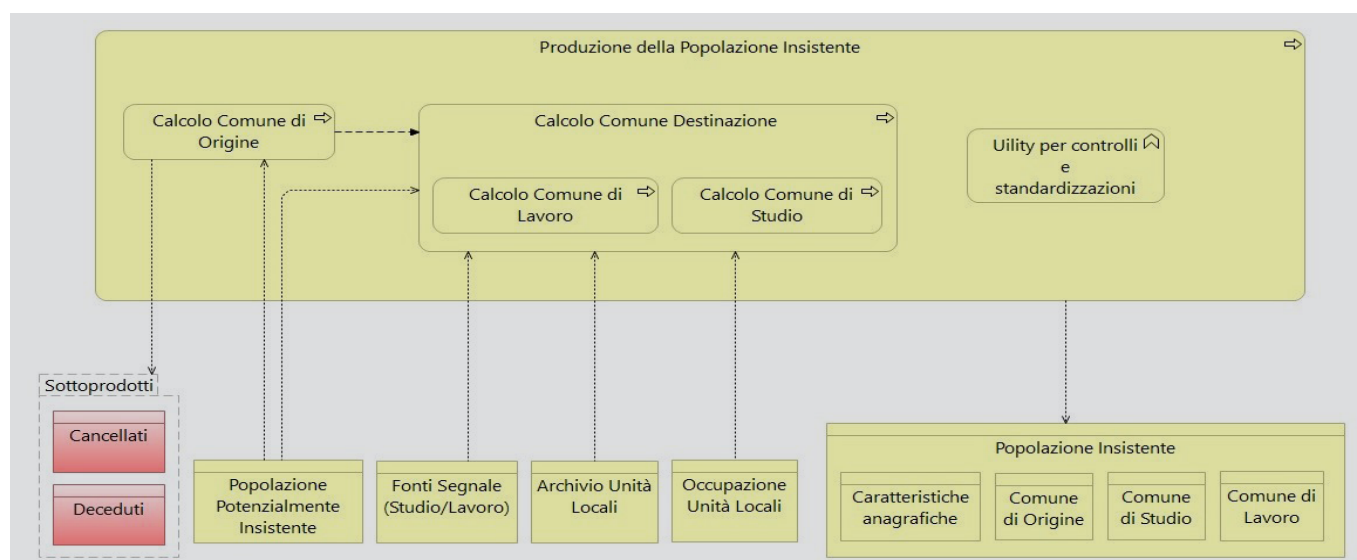
2. Base integrata “Popolazione insistente giornaliera per studio e lavoro”

Figura 2.3 - Schema del processo Caricamento da Registro delle Imprese attive e da Unità locali



Fonte: Elaborazione degli autori

Figura 2.4 - Schema del processo Produzione della Popolazione insistente



Fonte: Elaborazione degli autori

2.4 Mapping tra contenuti e definizioni della base “Popolazione insistente giornaliera per studio e lavoro - Archimede” e quelle del Registro base degli individui (RBI)

In parallelo alle attività di condivisione delle informazioni (su base e processi) e alla redazione dell'*as-is*, sono state condotte delle analisi per valutare la sostituibilità di alcuni input, e contestualmente un nuovo disegno per parti di processo, con le informazioni disponibili in RBI.

Sono state fatte delle valutazioni su:

- confrontabilità delle popolazioni di riferimento (*mapping* delle unità eleggibili);
- coerenza tra le classificazioni e i riferimenti temporali adottati per le variabili della base rispetto a quelle in RBI;
- analisi dei tracciati per individuare le variabili mutuabili da RBI.

L'esito di tali valutazioni ha permesso di individuare come obiettivo l'intervento sul processo Caricamento dati, fase Caricamento da fonti SIM, operando una sostituzione delle variabili anagrafiche e del luogo di origine con le corrispondenti variabili *core* in RBI (Prospetto 2.1).

Prospetto 2.1 - Elenco delle variabili utilizzate in “Popolazione insistente giornaliera per studio e lavoro - Archimede”, fonti di provenienza e corrispondenza con le variabili *core* in Registro base degli individui

ETICHETTA DELLA VARIABILE NELL'OUTPUT INTERMEDIO DELLA BASE	FONTI DI PROVENIENZA	ETICHETTA DELLA VARIABILE NEL REGISTRO
Codice_Fam	Lac	Cod_Famiglia
Codice_Provincia	Lac	Cod_Prov_Residenza
Codice_Comune	Lac	Cod_Com_Residenza
Data_Nascita	At se presente altrimenti moda in Sim_Individui	Data_Nascita
Sesso	At se presente altrimenti moda in Sim_Individui	Sesso
Prov_Nascita	At se presente altrimenti moda in Sim_Individui	Prov_Nascita
Com_Nascita	At se presente altrimenti moda in Sim_Individui	Com_nascita
Paese_Nascita	At se presente altrimenti moda in Sim_Individui	Paese_nascita
Data_Decesso	At o Cancellazioni Anagrafiche per morte	Data_decesso
Cittadinanza	Lac	Cod_cittadinanza

Fonte: Elaborazione degli autori

La valutazione dell'impatto potenziale di tale sostituzione è stata effettuata in modo indiretto tramite una analisi di concordanza tra variabili della base con variabili da registro (Paragrafo 2.6).

2.5 Alcune analisi di qualità: confronti con le statistiche ufficiali

La base dati preliminare è sottoposta ad un processo di controllo al fine di individuare le situazioni di incoerenza al suo interno.

Sulla base dati vengono sistematicamente effettuati controlli sui microdati (*missing*, fuori dominio, incoerenze tra variabili, ecc.) e controlli aggregati rispetto a fonti ufficiali di riferimento sintetizzati in indici di errore. Tali controlli permettono di mettere in luce e quantificare le incoerenze presenti, gli errori di classificazione, di misurazione, di rappresentazione (sovra-copertura e/o sotto-copertura della popolazione di riferimento) e valutare i valori estremi e anomali. Gli oggetti da controllare e le fonti di riferimento utilizzate sono elencate nel Prospetto 2.2.

2. Base integrata “Popolazione insistente giornaliera per studio e lavoro”

Prospetto 2.2 - Elenco oggetti presenti nella base e relative analisi

OGGETTO NELLA BASE	FONTI DI RIFERIMENTO	ANALISI EFFETTUATE
Comuni	Istat, Classificazione ufficiale dei Comuni	Numero comuni presenti nei dati di input coerenti con la classificazione ufficiale al 31.12.2017
Popolazione residente	Istat, Rilevazione annuale del movimento e calcolo della popolazione residente e della popolazione residente straniera	Differenze tra popolazione residente per sesso e classe di età e popolazione residente per comune nella base e nei Bilanci Demografici al 31.12.2017
Studenti scuole primarie/secondarie	Miur, Anagrafe nazionale degli studenti delle scuole	Differenze tra totale studenti nella base e nella fonte Miur
Studenti universitari	Miur, Anagrafe nazionale degli studenti universitari - Iscritti e iscrizioni	Differenze tra totale studenti universitari nella base e nella fonte Miur
Lavoratori	Istat, Rilevazione sulle forze di lavoro	Differenze percentuali tra lavoratori nella base e lavoratori nell'indagine Rfl (quarto trimestre) per provincia

Fonte: Elaborazione degli autori

Quelli qui presentati sono una parte dei controlli di qualità effettuati e riguardano, principalmente, la potenziale sotto-copertura delle popolazioni sul territorio legata al disallineamento delle diverse liste di comuni utilizzate istituzionalmente in Istituto e dagli Enti che forniscono i dati amministrativi. L'unione di informazioni territoriali da archivi amministrativi di diversi gestori, infatti, necessiterebbe di un lavoro di armonizzazione rispetto all'elenco dei comuni al 31 dicembre 2017³. I controlli (Tavola 2.1) mettono in luce e quantificano la presenza di disallineamento tra i comuni presenti negli archivi di input – e quindi nella base dati integrata intermedia riferita a dicembre 2017 – e la classificazione vigente.

Tavola 2.1 - Comuni italiani per tipo di relazione dell'individuo con il comune e validità rispetto alla classificazione ufficiale dei comuni italiani al 31 dicembre 2017 (valori assoluti)

COMUNI ITALIANI	Comuni presenti nei dati di input	di cui non validi al 31.12.2017	Individui nei comuni non validi
Comuni di origine	8.255	285	124.758
- Comune di residenza (da Lac)	7.954	13	33.087
- Comune di domicilio fiscale (da At)	8.255	285	91.671
Comune di studio (da Miur)	6.901	108	123.346
Comune di lavoro (da Inpdap, Inail, ...)	8.236	262	132.835
Comune di iscrizione universitaria (Miur)	266	0	0

Fonte: Elaborazione degli autori

Per la produzione della base 2017 sono state acquisite le Liste anagrafiche comunali (Lac) disponibili al 1 gennaio 2018: 7.954 comuni su 7.978 comuni ufficiali al 31 dicembre 2017. Il disallineamento è interamente dovuto alla fattispecie “cessazione del codice per fusione di diversi comuni o cambiamento di provincia dei comuni stessi”. Nel dettaglio, i tredici codici comune di residenza assegnati da Lac nella base intermedia e non presenti nella lista dei comuni ufficiali al 31 dicembre 2017, corrispondono ai trenta codici che sono invece presenti tra i comuni ufficiali al 31 dicembre 2017 e non nelle Lac. La differenza è, quindi, solo apparente ed è attribuibile alla diversa data di registrazione dell'elenco comuni nelle due fonti.

Per i comuni fuori classificazione si è optato per un intervento conservativo di pulizia del dato individuale, piuttosto che di correzione: i comuni non presenti nella lista ufficiale dei comuni al 31 dicembre 2017, per un totale di 33.087 individui, sono stati mantenuti nella base con un codice convenzionale di anomalia al posto del codice territoriale.

3 Elenco dei comuni al 31 dicembre 2017 (<http://www.istat.it/it/archivio/6789>).

La popolazione residente nella Rilevazione Istat annuale del movimento e calcolo della popolazione residente e della popolazione residente straniera al 31 dicembre 2017, pari a 60,484 milioni, è il riferimento rispetto al quale è stato effettuato il controllo della popolazione residente nella base dati “Popolazione insistente giornaliera per studio e lavoro - Archimede”. In particolare, è stato calcolato l'errore assoluto medio (MAPE⁴, *mean absolute percentage error*), pari alla media dei valori assoluti delle differenze relative tra la distribuzione della popolazione residente per comune della base rispetto alla distribuzione della popolazione residente per comune della rilevazione Istat (Tavola 2.2).

Tavola 2.2 - Differenze tra la popolazione residente nella base “Popolazione insistente giornaliera per studio e lavoro - Archimede” rispetto alla Rilevazione Istat al 31 dicembre 2017

MAPE - Media valore assoluto delle differenze percentuali 0,43

(*) Media dei valori assoluti delle differenze relative tra distribuzioni per comune

Fonte: Elaborazione degli autori

Valutando la distribuzione per ampiezza dell'errore, risulta che il 90,2 per cento dei comuni differisce per meno dell'1 per cento (Tavola 2.3).

Tavola 2.3 - Comuni per classe di differenza relativa tra popolazione residente nella base “Popolazione insistente giornaliera per studio e lavoro - Archimede” rispetto alla Rilevazione Istat al 31 dicembre 2017 (valori assoluti e composizione percentuale)

CLASSE DI DIFFERENZA RELATIVA %	N. comuni	% comuni	% cumulata comuni
0	877	11,0	11,0
]0-0,1]	1.558	19,6	30,7
]0,1-0,5]	3.614	45,5	76,2
]0,5-1]	1.115	14,0	90,2
>1	776	9,8	100,0
Totale	7.940	100,0	

Fonte: Elaborazione degli autori

La Popolazione straniera, nella Rilevazione annuale del movimento e calcolo della popolazione residente e della popolazione residente straniera - Istat, ammonta a 5,144 milioni. Nella base dati integrata è pari a 5,151 milioni. Anche in questo caso è stato calcolato l'errore medio (Tavola 2.4).

Tavola 2.4 - Differenze tra la popolazione residente di cittadinanza non italiana in “Popolazione insistente giornaliera per studio e lavoro - Archimede” rispetto alla Rilevazione Istat al 31 dicembre 2017

MAPE 3,91

(*) Media dei valori assoluti delle differenze relative tra distribuzioni per comune

Fonte: Elaborazione degli autori

Valutando la distribuzione per ampiezza dell'errore, si registra che il 91,5 per cento dei comuni differisce per meno dell'1 per cento (Tavola 2.5).

4
$$MAPE = \frac{\sum_{i=1}^N \left| \frac{a_i - b_i}{b_i} \right|}{N} \cdot 100$$

2. Base integrata “Popolazione insistente giornaliera per studio e lavoro”

Tavola 2.5 - Comuni per classe di differenza relativa tra popolazione residente di cittadinanza non italiana in “Popolazione insistente giornaliera per studio e lavoro - Archimede” rispetto alla Rilevazione Istat al 31 dicembre 2017 (valori assoluti e composizione percentuale)

CLASSE DI DIFFERENZA RELATIVA %	N. comuni	% comuni	% cumulata comuni
0	1.917	24,2	24,2
]0-0,1]	1.675	21,2	45,4
]0,1-0,5]	2.849	36,0	81,5
]0,5-1]	790	10,0	91,5
>1	673	8,5	100,0
Totale	7.906	100,0	

Fonte: Elaborazione degli autori

2.6 Confronto tra informazioni contenute nella base e nel Registro base degli Individui (RBI) al 31 dicembre 2017

In questa sezione vengono illustrati i principali risultati derivanti dai controlli sui microdati integrati della base con RBI. In particolare, nel Paragrafo 2.6.1 viene descritto il livello di sovrapposizione in termini di individui residenti presenti tra le due fonti e analizzate le caratteristiche dei “residui”. I Paragrafi 2.6.2 e 2.6.3 invece riportano, rispettivamente, il confronto tra le distribuzioni a livello territoriale subnazionale e gli esiti delle analisi di concordanza tra variabili individuali.

2.6.1 Concordanza rispetto alle unità

La Tavola 2.6 riporta le numerosità degli individui presenti in entrambe le fonti e dei residui. Si osserva un ottimo livello di sovrapposizione relativo agli individui residenti in vita al 31 dicembre 2017. Infatti, gli individui in comune tra le due fonti sono 60,4 milioni, pari al 99,8 per cento del totale degli individui in vita e residenti secondo quanto riportato in RBI al 31 dicembre 2017.

Tavola 2.6 - Individui residenti per presenza nelle due fonti. “Popolazione insistente giornaliera per studio e lavoro - Archimede” e Registro base degli individui. Anno 2017 (valori assoluti e composizione percentuale)

Individui in comune alle due fonti	60.400.234
% su totale individui residenti in RBI	99,8
% su totale individui residenti nella base	99,8
Individui presenti solo in RBI	
% su totale individui residenti in RBI	0,2
Individui presenti solo nella base	
% su totale individui nella base	0,2
Totale individui residenti nella base	60.524.298

Fonte: Elaborazione degli autori

2.6.2 Confronti tra le distribuzioni degli individui a livello territoriale subnazionale

Analogamente al precedente confronto con il riferimento Bilancio demografico, le distribuzioni degli individui per comune e provincia nelle due basi vengono confrontate in base al valore assunto dalla media delle differenze relative percentuali in valore assoluto (MAPE). Questo valore indica l'errore medio con cui i valori della base dati si discostano dai valori di RBI, che rappresentano, pertanto, quelli di riferimento. Anche il calcolo dell'indica-



tore è stato preceduto da verifiche sui codici comuni mancanti, variazioni e disallineamenti, procedendo al calcolo solo sui comuni presenti in entrambe le basi.

La Tavola 2.7 riporta i confronti relativi alla numerosità degli individui residenti nelle due fonti a livello territoriale provinciale e comunale.

Si osserva un buon livello di concordanza tra le due fonti: l'errore medio è pari allo 0,11 per cento a livello provinciale e 0,18 per cento per la distribuzione comunale.

Tavola 2.7 - Confronto tra distribuzioni territoriali nelle due fonti e medie delle differenze relative percentuali. "Popolazione insistente giornaliera per studio e lavoro - Archimede" e Registro base degli individui. Anno 2017

	Livello territoriale	MAPE
POPOLAZIONE RESIDENTE	Provinciale	0,11
	Comunale (a)	0,18

Fonte: Elaborazione degli autori

(a) Tutte le distribuzioni a livello comunale sono effettuate non considerando i comuni di Paderna (AL), Borgo Priolo (PV), Roccarainola (NA), Tufo (AV), Gallipoli (LE) e San Vito Lo Capo (TP), non presenti su PP nel senso che non hanno la Lac valorizzata.

2.6.3 Concordanza rispetto alle variabili core del Sistema integrato dei registri (SIR)

In questa sezione vengono riportati i risultati delle analisi condotte al fine di valutare la coerenza tra le due fonti, a livello individuale, relativamente alle variabili *core*. Le analisi sono riferite al solo sottoinsieme di individui in comune alle due fonti.

Rispetto alla variabile Data di nascita il grado di concordanza tra le due fonti è pari al 99,97 per cento (Tavola 2.8), mentre si hanno valori discordanti per circa 19 mila unità. Considerando le discordanze relative al solo anno di nascita si scende a meno di un migliaio di casi discordi (Tavola 2.9).

Tavola 2.8 - Individui per concordanza della variabile Data di nascita (ggmmyy). "Popolazione insistente giornaliera per studio e lavoro - Archimede" e Registro base degli individui. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA DATA DI NASCITA	N	%
Stessa data di nascita	60.380.930	99,97
Diversa data di nascita	19.304	0,03
Totale individui in comune	60.400.234	100,00

Fonte: Elaborazione degli autori

Tavola 2.9 - Individui per concordanza della variabile Anno di nascita (yy). "Popolazione insistente giornaliera per studio e lavoro - Archimede" e Registro base degli individui. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA ANNO DI NASCITA	N	%
Stesso anno di nascita	60.399.290	99,998
Diverso anno di nascita	944	0,002
Totale individui in comune	60.400.234	100,000

Fonte: Elaborazione degli autori

Se si considerano i singoli codici di cittadinanza secondo la classificazione degli stati esteri Istat in entrambe le fonti, si osserva che il 99,97 per cento degli individui presenta la stessa cittadinanza. Si registra, inoltre, una discordanza per 19,2 mila individui (Tavola 2.10).

2. Base integrata “Popolazione insistente giornaliera per studio e lavoro”

Tavola 2.10 - Individui per concordanza della variabile Cittadinanza. “Popolazione insistente giornaliera per studio e lavoro - Archimede” e Registro base degli individui. Anno 2017 (valori assoluti e composizione percentuale)

CONCORDANZA CITTADINANZA	N	%
Stesso codice cittadinanza	60.381.052	99,97
Codice cittadinanza diverso (a)	19.182	0,03
Totale individui in comune	60.400.234	100,00

Fonte: Elaborazione degli autori

(a) Sono compresi anche i nulli nella base.

Analizzando nel dettaglio i casi per cui il codice cittadinanza della base non coincide con quello di RBI, si osserva che si tratta, in prevalenza, di individui per cui la cittadinanza presente sulla base è riferita a codici di cittadinanza non validi, ovvero non inclusi nell'elenco dei codici degli stati esteri ufficiale. Si tratta di codici riferiti a stati che hanno subito delle variazioni o cessazioni nel tempo, oppure di errori di codifica presenti nella fonte di input.

2.7 Prime conclusioni e sviluppi futuri

La fase di definizione dell'*as-is* è stata un'occasione di analisi approfondita dei processi in essere. Tale analisi ha messo in evidenza che, soprattutto nella fase di acquisizione e pre-trattamento delle fonti, sarebbe necessaria una revisione del processo (ad es. per l'eliminazione di variabili non più utilizzate o ridondanti). Il processo di armonizzazione della base nel SIR potrebbe, quindi, essere l'occasione per rendere l'intero processo più efficiente e, di conseguenza, il rilascio della base più tempestivo, indipendentemente dalla sostituzione di contenuti informativi della base con quelli dei registri.

Per quanto riguarda, invece, la sostituzione con le variabili acquisite dalle fonti, i controlli effettuati per valutare il grado di coerenza tra la base e RBI hanno mostrato un elevato grado di coerenza tra le variabili nelle due fonti. Ci si attende, quindi, che la sostituzione delle variabili della base con quelle di RBI non avrà impatto sulle serie storiche degli indicatori prodotti. Oltre a ciò, per alcune variabili (come la cittadinanza) la sostituzione dell'informazione della base con quella di RBI apporterà un incremento di qualità.

Il lavoro fin qui svolto ha condotto all'individuazione di una prima serie di modifiche e sostituzioni di alcune variabili con quelle contenute in RBI. Tuttavia, il sistema dei registri si configura come un sistema in evoluzione continua. Pertanto, è possibile prevedere una modellazione “aperta” ad altri registri. In particolare, si dovranno effettuare più approfondite analisi per valutare se sarà possibile, in futuro, sostituire i segnali di studio/lavoro e i rispettivi luoghi di lavoro/studio con informazioni provenienti dai registri.

3. BASE INTEGRATA “PRECARIETÀ LAVORATIVA”¹

3.1 Premessa

Il presente capitolo ha l'obiettivo di illustrare lo stato di avanzamento dei lavori della base “Precarietà lavorativa - Archimede” (base PL). Nella prima parte del documento si fornisce una descrizione sintetica della base (Paragrafo 3.2). Vengono poi esposti i relativi contenuti informativi della base dati (Paragrafo 3.3), il *mapping* rispetto al Registro statistico di base degli individui, delle famiglie e delle convivenze (RBI) e la modellazione del processo (Paragrafo 3.4). Successivamente, si presentano alcuni risultati relativi a misure di qualità della base rispetto alle statistiche ufficiali (Paragrafo 3.5) e parte dei risultati del confronto tra le variabili comuni alla base ed RBI (Paragrafo 3.6). Il paragrafo su conclusioni e sviluppi futuri conclude il documento (Paragrafo 3.7).

3.2 Contesto

La base dati “Precarietà lavorativa - Archimede” ha lo scopo di analizzare l'universo dei lavoratori precari, all'interno di quello degli occupati regolari, descrivendone gli aspetti demo-sociali ed economico-lavorativi sia in un determinato periodo sia seguendone le trasformazioni nel tempo.

La popolazione di riferimento è dunque costituita dagli individui che nel periodo di riferimento sono titolari di uno o più contratti di lavoro regolari, anche svolti contemporaneamente.

Gli occupati regolari sono osservati nel mese di ottobre dell'anno e sono quelli che in tale mese hanno un segnale di lavoro in specifiche fonti amministrative selezionate in funzione di obiettivi di qualità, copertura dell'universo, tempestività. Le fonti sono state analizzate al fine di estrarre, elaborare e ottenere il seguente output: una base di dati corrispondenti ai singoli lavoratori, rappresentati dall'attività lavorativa prevalente svolta durante il periodo di riferimento dello studio. Nell'integrazione delle fonti è stata utilizzata una chiave composta da due variabili “Lavoratore - Attività lavorativa prevalente”, dove, nel caso di molteplicità di attività di lavoro nel periodo di riferimento, la prevalente è individuata in base ad un ordinamento per stabilità lavorativa.

Le fonti amministrative selezionate che sono alla base del processo di integrazione hanno natura fiscale e previdenziale e sono le seguenti:

- la fonte Istat - Registro statistico dell'occupazione delle imprese (Asia Occupazione), che include anche l'archivio Inps - Gestione dipendenti pubblici (ex Inpdap) - Posizioni Assicurative, è il principale contenuto informativo;
- la fonte Inps - Lavoratori agricoli autonomi, che riguarda gli imprenditori agricoli (coltivatori diretti e imprenditori agricoli professionali), ossia coloro che esercitano un'attività diretta alla coltivazione del fondo, alla silvicoltura, all'allevamento di animali e attività connesse;

¹ I Paragrafi 3.1, 3.2, 3.3 e 3.7 sono a cura di Dario Ercolani; il Paragrafo 3.4 è a cura di Adele Maria Bianco; i Paragrafi 3.5 e 3.6 sono a cura di Fulvia Cerroni.

- la fonte Inps - Rapporto di lavoro domestico, che contiene informazioni su coloro che prestano un'attività lavorativa continuativa per le necessità della vita familiare del datore di lavoro, come ad esempio colf, assistenti familiari o babysitter, governanti, camerieri, cuochi ed altri operatori connessi;
- le Liste Anagrafiche Comunali (Lac), che rappresentano l'elenco depositato e aggiornato presso ogni comune italiano dei residenti sul territorio di competenza, delle loro caratteristiche anagrafiche e della dimensione e tipologia familiare di appartenenza.

3.3 Contenuti informativi della base dati e definizioni principali

La costruzione della collezione di dati ha tenuto conto di numerosi aspetti connessi al contenuto e alla qualità delle informazioni incluse nelle fonti amministrative utilizzate. A valle delle analisi di qualità delle singole fonti, sono state effettuate attività di integrazione, omogeneizzazione e trasformazione.

Le principali caratteristiche lavorative di interesse ricostruite per il mese di ottobre sono: numero di datori di lavoro; datore di lavoro principale; attività lavorativa prevalente; condizione prevalente; condizione prevalente nell'anno precedente; intensità lavorativa mensile dell'attività prevalente; intensità lavorativa mensile complessiva; presenza di segnale di Cassa integrazione guadagni e/o Contratto di solidarietà. I principali aspetti definitori richiamati nella costruzione della base in questione sono:

- *Lavoratore precario*: individuo che nel periodo di riferimento svolge principalmente lavori di tipologia contrattuale instabile (dipendente a tempo determinato, collaboratore, lavoro domestico);
- *Intensità lavorativa*: quota di periodo di riferimento lavorata rispetto al periodo lavorabile (il mese).

3.4 Modellazione del processo

Il processo attuale è stato descritto mediante diagrammi Archimate, linguaggio di modellazione che supporta la descrizione, l'analisi e la visualizzazione dell'architettura di un processo.

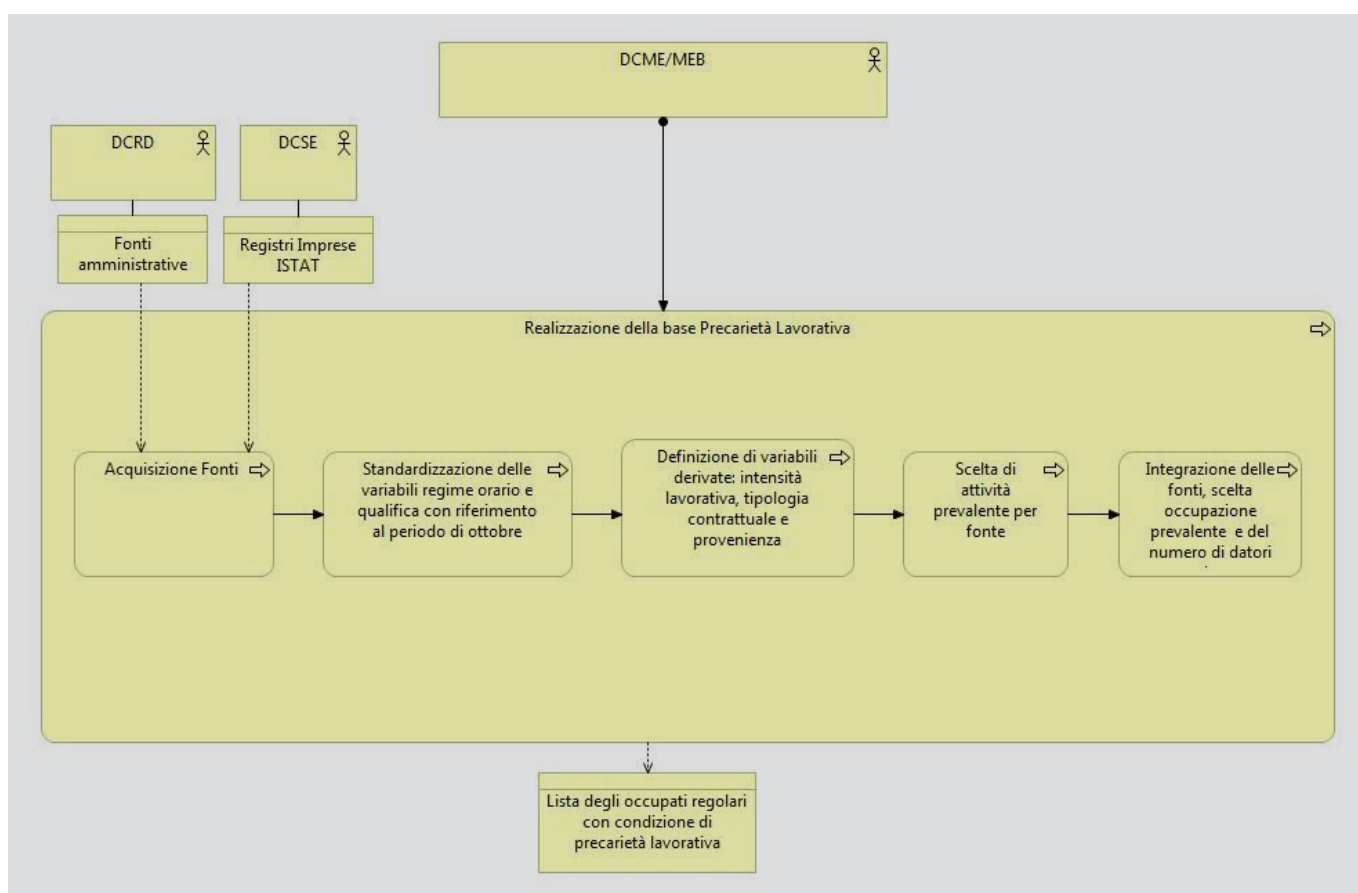
Al processo prendono parte tre attori: la Direzione centrale per la metodologia ed il disegno dei processi statistici (DCME) per la metodologia di design con il ruolo di responsabile dell'intero processo, la Direzione centrale per la raccolta dati (DCRD) come responsabile delle fonti amministrative e la Direzione centrale per le statistiche economiche (DCSE) come responsabile dei registri Istat sulle imprese.

Nella Figura 3.1 viene riportata una visione globale del processo "Realizzazione della base Precarietà lavorativa - Archimede". Esso prende in input due gruppi di dati: alcuni derivanti da fonti amministrative elementari acquisite direttamente dal Sistema integrato dei microdati (SIM) e altre da registri statistici Istat (es. archivi d'impresa); in output viene fornita la lista degli occupati regolari classificati secondo la condizione prevalente (stabile o precaria). Il processo può essere a sua volta suddiviso in cinque fasi (dettagliate nella Figura 3.2).

3. Base integrata “Precarietà lavorativa”

Nella fase di acquisizione delle fonti vengono eseguiti due sotto-processi: il primo riguarda la selezione delle variabili del datore di lavoro e del lavoratore dalle fonti amministrative Inps - Rapporti di lavoro domestico, Inps - Lavoratori agricoli autonomi e il Registro Asia Occupazione. Completato questo primo sotto-processo, si passa alla standardizzazione di un sottoinsieme delle variabili secondo le classificazioni ad esse associate. A questo punto del processo vengono derivate nuove variabili (fase 3), nel caso di qualitative con le relative classificazioni (es. tipologia contrattuale).

Figura 3.1 - As-is schema generale del processo di produzione relativo alla base dati “Precarietà lavorativa – Archimede”



Fonte: Elaborazione degli autori

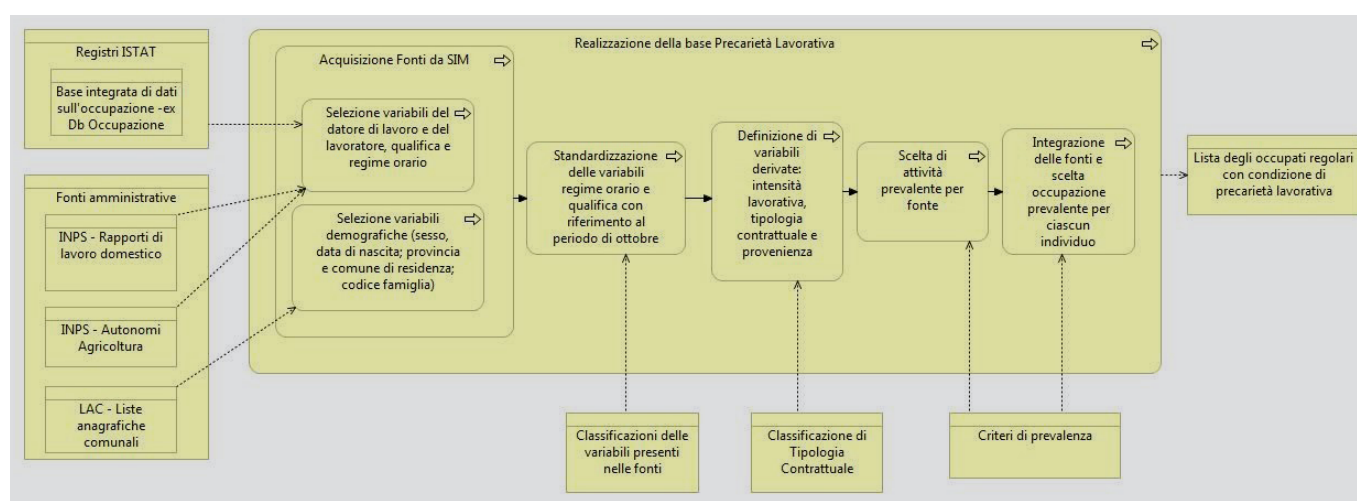
Infine si procede con la scelta dell'attività prevalente, inizialmente per ogni singola fonte e successivamente tra tutte le fonti utilizzando una lista di criteri di prevalenza. Inoltre, per ciascun occupato viene elaborata la variabile del numero di datori di lavoro.

I criteri di individuazione dell'attività lavorativa prevalente recepiscono le indicazioni dell'Indagine censuaria sulla popolazione e della Rilevazione sulle forze di lavoro, che a loro volta seguono i Regolamenti Europei.

In sintesi, gli elementi considerati che sono stati usati al fine di attribuire l'attività lavorativa prevalente all'individuo occupato della base "Precarietà lavorativa - Archimede" sono, in ordine di applicazione:

1. il periodo di lavoro osservabile (periodo di versamento contributivo; periodo di iscrizione in archivio; è comunque una misura dipendente dal contenuto informativo della fonte utilizzata);
2. stabilità dell'attività lavorativa (in termini di scadenza contrattuale, debolezza contrattuale o dipendenza economica dai propri datori di lavoro).

Figura 3.2 - Dettaglio delle fasi del processo di produzione relativo alla base dati "Precarietà lavorativa - Archimede"



Fonte: Elaborazione degli autori

3.4.1 Mapping tra contenuti della base Precarietà lavorativa e quelle del Registro base degli Individui (RBI) e definizioni

L'impatto della coerenza dei contenuti informativi della base dati "Precarietà lavorativa - Archimede" con le analoghe informazioni derivanti da RBI è stata preceduta da una analisi sulle variabili contenute in entrambe le fonti. Il prospetto seguente riporta il risultato di tale mappatura con riferimento all'annualità 2017.

Prospetto 3.1 - Mapping delle principali variabili di "Precarietà lavorativa - Archimede" nel Registro base degli Individui (RBI). Anno 2017

TIPO DI VARIABILE	DESCRIZIONE VARIABILE	ETICHETTA DELLA VARIABILE IN PRECARIETA' LAVORATIVA - ARCHIMEDE	FONTI DELLA VARIABILE IN PRECARIETA' LAVORATIVA - ARCHIMEDE	ETICHETTA DELLA VARIABILE IN RBI
Variabili individuali anagrafiche	Sesso	SESSO	Liste anagrafiche comunali	SESSO
	Data di nascita	ANNO		DATA_NASCITA
Variabili individuali di residenza anagrafica	Codice provincia di residenza	COD_PRO	Liste anagrafiche comunali	COD_PROV_RESIDENZA
	Codice comune di residenza	COD_COM		COD_COM_RESIDENZA

Fonte: Elaborazione degli autori

3.5 Qualità rispetto alle statistiche ufficiali

Di seguito è riportata un'analisi della qualità della fonte di dati "Precarietà lavorativa - Archimede" messa a confronto prima con le statistiche ufficiali Istat prodotte dalla Rilevazione sulle forze di lavoro e in seconda battuta con RBI (Paragrafo 3.6).

3. Base integrata “Precarietà lavorativa”

Nel prospetto seguente si riportano i risultati del primo dei due confronti effettuato sotto il profilo dei metadati, mettendone in luce le differenze esistenti a partire dai metadati concettuali (obiettivi, aspetti definitori e popolazione di riferimento) per passare poi ai metadati metodologici (tipologia di fonte e periodo di riferimento) arrivando poi a studiare i metadati relativi alla qualità. Tali contenuti infatti descrivono la qualità delle statistiche declinandola secondo le diverse dimensioni riconosciute a livello internazionale (ad esempio tempestività e accuratezza). La qualità viene qui principalmente esplosa sotto il profilo dell'accuratezza.

Prospetto 3.2 - Confronto tra le fonti Rilevazione sulle Forze di Lavoro e Precarietà lavorativa - Archimede

	Forze di Lavoro	vs	Progetto ARCHIMEDE - PL
•Obiettivi	Mercato del lavoro		Precarietà lavorativa
•Aspetti definitori	Lavoratore «atipico»		Lavoratore «precario»
•Popolazione di riferimento	Famiglie residenti		Occupati presenti negli Archivi
•Tipologia fonte	Campionaria		Fonti amministrative
•Periodo di riferimento	Settimana		Dettaglio connesso alla fonte

Fonte: Elaborazione degli autori

Opportunamente riclassificata per condizione professionale (dipendente/indipendente) e carattere del lavoro indipendente (autonomo-collaboratore), la distribuzione di “Precarietà lavorativa - Archimede”, che è riferita al mese di ottobre, risulta confrontabile con quella della Rilevazione sulle forze di lavoro (Rfl) riferita al IV trimestre dell'annualità del 2017 (Tavola 3.1). Dal confronto emerge una differenza percentuale calcolata tra i totali dei lavoratori occupati di appena lo 0,1 per cento; disaggregando per condizione professionale si evidenzia sulla base una sottostima del lavoro dipendente (-2,3 per cento) e una sovrastima di quello indipendente (2,3 per cento). Questo spiega il fatto che l'errore assoluto percentuale tra gli aggregati totali sia tendente a zero, come risultato di un effetto compensatorio della sottostima del lavoro dipendente e della sovrastima di quello indipendente. Il MAPE (*Mean Absolute Percentage Error*) sulla stessa variabile a livello Italia, calcolato come media tra gli errori sui dipendenti (2,9 per cento) e sugli indipendenti (10,3 per cento), è pari a 6,6 per cento.

Tavola 3.1 - Occupati per condizione professionale. “Precarietà lavorativa - Archimede” e Rilevazione sulle forze di lavoro IV trimestre. Anno 2017 (valori assoluti, dati in migliaia)

CONDIZIONE PROFESSIONALE	Precarietà lavorativa - Archimede		Rilevazione sulle forze di lavoro		Differenza tra %	Absolut Percentage Error
	Valori assoluti	Valori %	Valori assoluti	Valori %		
Dipendenti	17.270	74,7	17.791	77,1	-2,3	2,9
Indipendenti	5.846	25,3	5.298	22,9	2,3	10,3
di cui						
- Autonomi	5.631	24,4	5.040	21,8	2,5	11,7
- Collaboratori	214	0,9	259	1,1	-0,2	17,2
Totale	23.115	100	23.090	100	0,0	0,1

Fonte: Elaborazione degli autori

Nella Tavola 3.2 il confronto sulla condizione professionale dei lavoratori tra “Precarietà lavorativa - Archimede” e Rfl viene declinato per ripartizione geografica, mostrando dal punto di vista territoriale una minor distanza dapprima tra le distribuzioni percentuali degli autonomi rispetto ai dipendenti, e poi tra i collaboratori. Il MAPE sulla ripartizione geografica – media degli errori sulle ripartizioni – è pari al 3,9 per cento.

Tavola 3.2 - Occupati per ripartizione geografica non missing e condizione professionale. “Precarietà lavorativa - Archimede” e Rilevazione sulle forze di lavoro IV trimestre. Anno 2017 (valori percentuali)

CONDIZIONE PROFESSIONALE	Precarietà lavorativa - Archimede				Rilevazione sulle forze di lavoro				Differenza tra %			Absolut Percentage Error
	CONDIZIONE PROFESSIONALE				CONDIZIONE PROFESSIONALE							
	Dipendenti	Autonomi	Collaboratori	Totale	Dipendenti	Autonomi	Collaboratori	Totale	Dipendenti	Autonomi	Collaboratori	
Nord-ovest	29	28	24	29	30	28	29	30	-1,0	-0,3	-4,9	4,8
Nord-est	22	22	17	22	22	21	18	22	0,0	0,7	-1,1	1,1
Centro	21	21	24	21	22	21	25	21	-1,0	0,0	-1,6	5,4
Sud e Isole	28	29	36	28	26	29	28	27	2,0	-0,3	7,6	4,1

Fonte: Elaborazione degli autori

Nella Tavola 3.3, il confronto tra “Precarietà lavorativa - Archimede” e Rfl viene esploso secondo il carattere dell’occupazione lavorativa alle dipendenze specificandone se si tratta di contratto a tempo determinato o indeterminato. Nella Rfl sono stati selezionati i dati grezzi (ovviamente i dati amministrativi non sono destagionalizzati) – quindi non corretti attraverso destagionalizzazione – riferiti al mese di ottobre 2017.

Tavola 3.3 - Occupati per condizione professionale e carattere dell’occupazione. “Precarietà lavorativa - Archimede” e Rilevazione sulle forze di lavoro (dati mensili grezzi). Mese ottobre Anno 2017 (valori assoluti, dati in migliaia)

CONDIZIONE PROFESSIONALE	CARATTERE CCUPAZIONE	Precarietà lavorativa - Archimede		Rilevazione sulle forze di lavoro		Differenza tra %	Absolut Percentage Error
		Valori assoluti	Valori %	Valori assoluti	Valori %		
Dipendenti	Tempo determinato	2.874	12,4	2.815	12,2	0,3	2,1
	Tempo indeterminato	14.396	62,3	15.022	64,9	-2,6	4,2
	Totale	17.270	74,7	17.837	77,1	-2,3	3,2
Indipendenti		5.846	25,3	5.312	22,9	2,3	10,0
Totale		23.115	100	23.150	100	-	0,1

Fonte: Elaborazione degli autori

3. Base integrata “Precarietà lavorativa”

La sottostima dei dipendenti presenti nella base rispetto alla Rfl (-2,3 per cento), già tracciata nella Tavola 3.1, qui si arricchisce dell'informazione sul carattere nell'occupazione evidenziando che tale sottostima si concentra tra i dipendenti a tempo indeterminato. L'errore medio sul carattere di occupazione (media tra tempo determinato e tempo indeterminato) è pari a 3,1 per cento.

Nella Tavola 3.4 il confronto tra “Precarietà lavorativa - Archimede” e la Rilevazione sulle forze di lavoro esplora la struttura anagrafica della popolazione occupata secondo le variabili sesso e classe di età.

Sia in valore assoluto sia in percentuale le distribuzioni confrontate mostrano una elevata vicinanza nel carattere sesso, e lo stesso si può dire per la distribuzione secondo le classi di età, ad esclusione però del confronto per le classi in età lavorativa più estreme (tranne quindi le classi d'età più lontane: 15-24 anni e più di 65).

Tavola 3.4 - Occupati per sesso e classi di età. “Precarietà lavorativa - Archimede” (con sesso e classi d'età non missing) e Rilevazione sulle forze di lavoro (dati mensili grezzi). Mese ottobre Anno 2017 (valori assoluti, dati in migliaia)

		Precarietà lavorativa - Archimede		Rilevazione sulle forze di lavoro		Absolut Percentage Error
		Valori assoluti	Valori %	Valori assoluti	Valori %	
SESSO	Femmine	9.794	42,4	9.743	42,1	0,5
	Maschi	13.300	57,6	13.407	57,9	0,8
CLASSI DI ETÀ'	15-24	1.074	4,6	995	4,3	7,9
	25-34	4.022	17,4	4.053	17,5	0,8
	35-49	9.418	40,8	9.838	42,5	4,3
	50-64	7.664	33,2	7.662	33,1	0,0
	65+	916	4,0	602	2,6	52,2

Fonte: Elaborazione degli autori

3.6 Confronto tra informazioni contenute nella base e nel Registro base degli Individui (RBI)

L'analisi prosegue mediante il confronto di “Precarietà lavorativa - Archimede” con le informazioni contenute nel RBI, ovvero dal punto di vista della copertura della fonte in termini di unità in comune e unità presenti solo nella base.

In Tavola 3.5 si osserva che solo l'1,8 per cento degli individui della base non è presente in RBI e sono tutti individui senza comune di residenza valido. Riguardo gli individui che hanno tutti il comune di residenza valido si può fare un confronto con l'informazione secondo il carattere sesso e secondo le classi di età contenuta in RBI (Tavole 3.6 e 3.7).

Tavola 3.5 - Individui per presenza nelle due fonti. “Precarietà lavorativa - Archimede” e Registro base degli individui. Anno 2017 (valori assoluti e percentuali, dati in migliaia)

Totale individui in Precarietà lavorativa - Archimede	23.115
Individui residenti con comune di residenza valorizzato	22.693
<i>Incidenza percentuale su tot. individui in Precarietà lavorativa - Archimede</i>	98,2%
Individui residenti in comune alle fonti Precarietà Lavorativa e RBI	22.693
<i>Incidenza percentuale su tot. individui residenti in Precarietà lavorativa - Archimede</i>	100%
<i>Incidenza percentuale su tot. individui residenti in RBI</i>	37,5%
Individui residenti presenti solo in PL	
<i>Incidenza percentuale su tot. Individui in Precarietà lavorativa - Archimede</i>	1,8%

Fonte: Elaborazione degli autori



Tavola 3.6 - Individui in comune nelle fonti “Precarietà lavorativa – Archimede” e Registro base degli individui per sesso. Anno 2017 (valori assoluti, dati in migliaia)

FONTE/SESSO	Registro base degli individui		
	Femmine	Maschi	Totale
Precarietà lavorativa-Archimede			
Femmine	9.674	-	9.674
Maschi	-	13.019	13.019
Totale	9.674	13.019	22.693

Fonte: Elaborazione degli autori

Tavola 3.7 - Individui in comune nelle fonti “Precarietà lavorativa – Archimede” e Registro base degli individui per classe di età. Anno 2017 (valori assoluti, dati in migliaia)

FONTE/CLASSI DI ETÀ	Registro base degli individui				
	15-24	25-34	35-49	50-64	65+
Precarietà lavorativa-Archimede					
15-24	1.042	-	-	-	-
25-34	-	3.928	-	-	-
35-49	-	-	9.250	-	-
50-64	-	-	-	7.573	-
65+	-	-	-	-	900
Totale	1.042	3.928	9.250	7.573	900

Fonte: Elaborazione degli autori

Il confronto rispetto alle distribuzioni per tipologia contrattuale riportato nella Tavola 3.8 mostra come le quote percentuali siano sovrapponibili e quindi non ci sarebbero perdite significative di informazione in nessuna delle tipologie contrattuali nel passare da tutta la popolazione della base a quella dei soli individui in comune con RBI, poiché essa è costituita da ben il 98,2 per cento della base.

Tavola 3.8 - Tutti gli individui della fonte “Precarietà lavorativa - Archimede” e solo gli individui in comune con Registro base degli individui (RBI) per tipologia contrattuale. Anno 2017 (valori assoluti, dati in migliaia)

TIPOLOGIA CONTRATTUALE	Precarietà lavorativa - Archimede		Precarietà lavorativa - Archimede in comune con RBI		Absolut Percentage Error
	Valori assoluti	Valori %	Valori assoluti	Valori %	
Dipendente a T.I. part time	3.084	13,3	3.031	13,4	1,7
Autonomo	5.631	24,4	5.497	24,2	2,4
Dipendente a T.D.	2.874	12,4	2.756	12,1	4,1
Collaboratore	214	0,9	209	0,9	2,5
Lavoratore domestico	586	2,5	553	2,4	5,6
Totale	23.115	100	22.693	100	1,8
Totale	23.115	100	23.090	100	0,1

Fonte: Elaborazione degli autori

3.7 Conclusioni e sviluppi futuri

L'analisi del processo con la disamina delle attività di realizzazione della base “Precarietà lavorativa - Archimede” ha messo in luce alcuni possibili miglioramenti connessi alla sostituzione della fonte costituita dalle Liste anagrafiche comunali con RBI, in virtù del fatto che quest'ultimo è un registro dalla struttura e contenuti più stabili nel tempo.

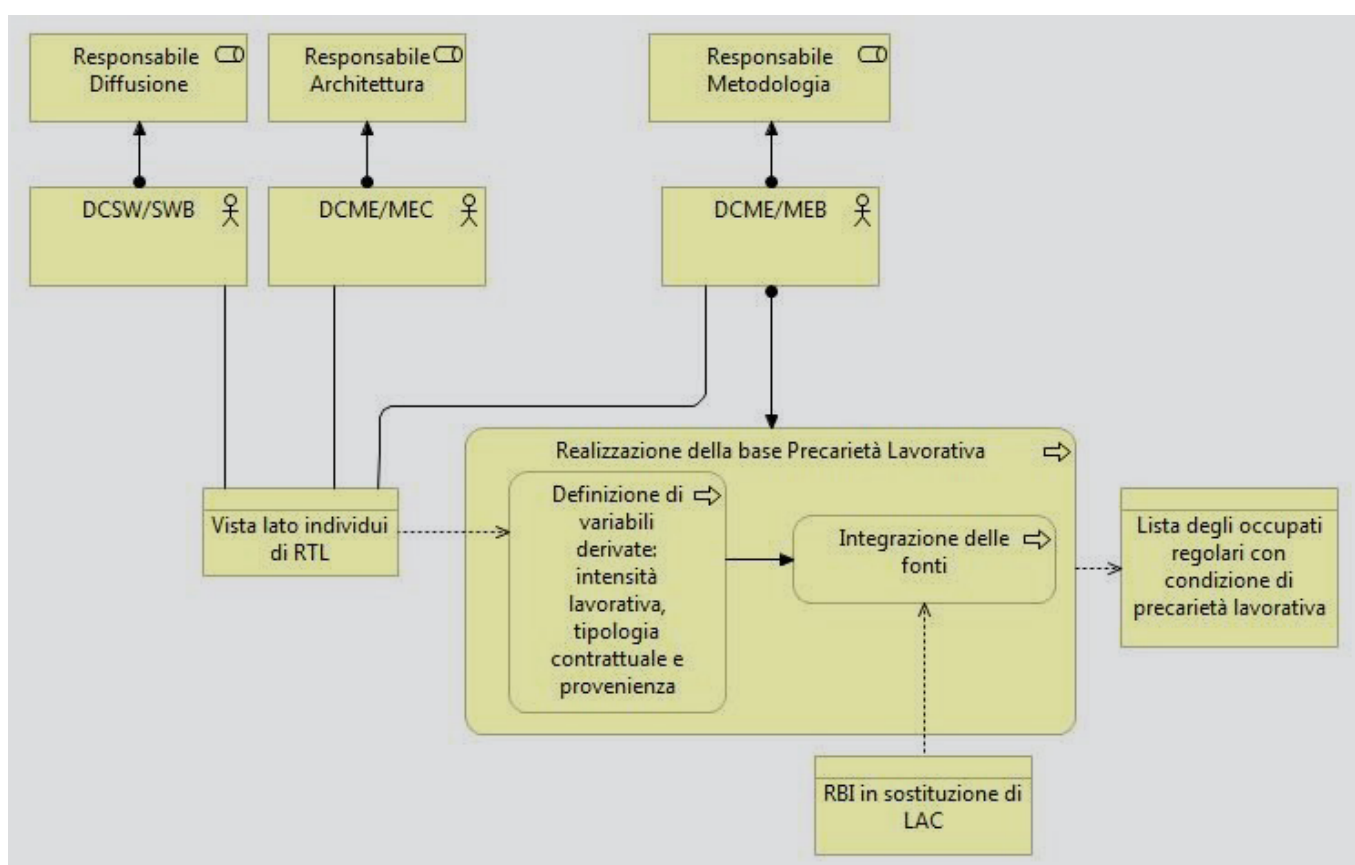
Inoltre, da un lato la sostituzione non influisce sull'ammontare della base, dall'altro non diminuisce la stabilità qualità delle serie storiche, nonché le distribuzioni dei principali

3. Base integrata “Precarietà lavorativa”

caratteri anagrafici sono ben approssimate tra loro. Ad avvalorare tale tesi è il fatto che la numerosità della base è molto simile a quella della fonte di riferimento primaria sul mercato del lavoro che è la Rilevazione sulle forze di lavoro.

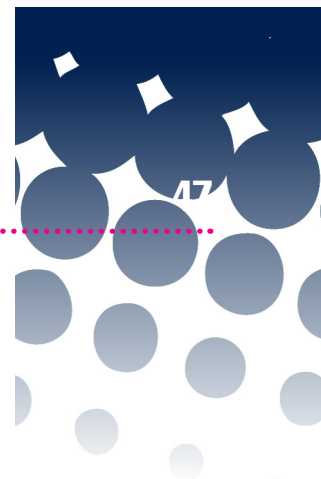
Riguardo un ipotetico sviluppo – oggetto di futuri studi – va affermato che le principali variabili di input e di rilascio della base sulla precarietà lavorativa riguardano il mercato del lavoro. Di conseguenza, di fondamentale importanza per l’armonizzazione della base nel SIR sarà l’utilizzo del Registro statistico tematico del lavoro (RTL). Nel diagramma che segue (Figura 3.3) viene riportata una possibile versione a tendere (*to-be*) del processo.

Figura 3.3 - Ipotesi di sviluppo futuro del processo di produzione relativo alla base dati “Precarietà lavorativa - Archimede”



Fonte: Elaborazione degli autori

In particolare: come detto, RBI potrebbe sostituire completamente l’alimentazione costituita allo stato attuale dalle Liste anagrafiche comunali. In prospettiva la Base degli individui occupati, che sarà derivata dal RTL, e che al momento è in fase di studio, potrebbe rimpiazzare le altre fonti amministrative relative al mercato del lavoro.



4. PROPOSTA DI UNA NUOVA ARCHITETTURA ARMONIZZATA CON IL SISTEMA INTEGRATO DEI REGISTRI (SIR) DELLE BASI DI ARCHIMEDE¹

4.1 Premessa

L'attività di analisi del processo attuale di produzione delle tre basi descritte nei Paragrafi 1.2, 2.3 e 3.2 (analisi *as-is*) ne ha reso possibile la modellazione dell'architettura integrata con il Sistema integrato dei registri (SIR). Nel progettarela sono stati delineati i processi di gestione dell'aggiornamento dell'alimentazione, dell'armonizzazione delle tre basi Archimede con il registro RBI e la definizione delle azioni da intraprendere per delineare futuri scenari di integrazione di Archimede con tutti gli altri registri del SIR.

Nei seguenti paragrafi saranno presentati la modellazione del processo di produzione delle basi Archimede dopo la sostituzione dell'alimentazione di alcune fonti con il Registro base degli individui e delle convivenze (RBI) e con il Registro statistico di base dei luoghi (RSBL), e la gestione della loro armonizzazione con il SIR (Paragrafo 4.2). La rappresentazione grafica dell'alimentazione di Archimede dopo l'armonizzazione con il SIR e alcuni possibili scenari evolutivi di tali integrazioni (Paragrafo 4.3). La valutazione di impatto della nuova architettura e la proposta di aggiornamento in diverse fasi (Paragrafo 4.4).

4.2 Modellazione del processo di produzione delle basi di Archimede armonizzato con il Sistema integrato dei registri (SIR)

La modellazione del processo che si potrà ottenere integrando Archimede con i registri del SIR deve garantire il mantenimento della qualità dei dati della attuale base Archimede e il miglioramento del livello di standardizzazione del processo e di armonizzazione con il principale sistema dei dati dell'istituto, il SIR. L'alimentazione dal SIR inoltre garantirà in futuro ad Archimede sia di godere del miglioramento della qualità dei dati di questo sistema, che dell'ampliamento dei dati integrati e quindi dell'offerta informativa del SIR.

L'analisi della nuova architettura (versione *to-be*) ha preso le mosse dai risultati delle analisi di qualità descritte nei primi tre capitoli, in particolare da quelle condotte sull'effetto della potenziale alimentazione da RBI. In funzione di queste sono state proposte evoluzioni future considerando l'ambito informativo delle basi Archimede e la possibilità attuale e futura di contributo a tale ambito da parte del SIR. Per queste nuove opportunità di integrazione è necessaria una attenta analisi di qualità.

Per quanto riguarda l'armonizzazione rispetto alle variabili territoriali sarà necessario allineare i dati con quelli del Registro RSBL, poiché contiene l'informazione relativa alla gerarchia del territorio: Provincia, Unità territoriali sovracomunali (UTS), Regioni e Ripartizioni. Inoltre questa integrazione garantirebbe la gestione del territorio nel tempo, renden-

¹ I Paragrafi 4.1, 4.3 e 4.4 sono a cura di Roberta Radini; il Paragrafo 4.2.1 è a cura di Domenico Infante; il Paragrafo 4.2.2 è a cura di Paolo Francescangeli; il Paragrafo 4.2.3 è a cura di Maria Adele Bianco.



do automatiche le trasformazioni amministrative del territorio. La gestione della coerenza spazio-temporale è demandata a questo registro che consente anche le operazioni di attualizzazione (trasformazione dei dati territoriali nel tempo in avanti rispetto al tempo di riferimento dei dati stessi), sia la de-attualizzazione (spostamento dei dati indietro nel tempo).

4.2.1 Modellazione del processo di produzione della Base Condizioni socio economiche delle famiglie - Archimede armonizzato con il SIR

Nel Capitolo 1 è stato già presentato il processo di produzione della base “Condizioni socio economiche delle famiglie - Archimede”, con una dettagliata rappresentazione degli attori coinvolti ed i ruoli ricoperti, delle fonti dati utilizzate; inoltre sono stati delineati i principali processi che portano alla creazione della base stessa.

Allo scopo di analizzare i guadagni di qualità ottenibili dalla sostituzione delle fonti amministrative con RBI, nel Paragrafo 1.4, è stata schematizzata la corrispondenza tra informazioni provenienti dai Registri (con dettaglio da RBI) e quelle della base sulle “Condizioni socio economiche delle famiglie”: tale analisi ha evidenziato un alto livello di concordanza tra le informazioni presenti nei sistemi sia al livello macro (dati aggregati) che micro (singoli record), questo ha portato alle conclusioni riportate nel Paragrafo 1.5, in merito alla possibile sostituibilità di alcuni dati da fonte, utilizzati dalla base, con informazioni provenienti da Registri.

Dalle predette conclusioni è stata avviata l’analisi della nuova architettura della base e quindi la descrizione del nuovo processo, evidenziando gli interventi di sostituzione delle fonti con dati da Registri in modo da fornire una visione prospettica di come potrebbe evolvere l’architettura del processo in esame (versione *to-be*). La sua modellazione evidenzierà i miglioramenti ottenibili sia a livello di semplificazione delle fonti necessarie alla costruzione della base “Condizioni socio economiche delle famiglie-Archimede”, sia nella semplificazione dei sotto-processi di elaborazione, sia nella qualità complessiva. Si evidenzia inoltre che il registro base RBI utilizza esclusivamente codici identificativi dell’individuo pseudo-anonimizzati e non codici identificativi diretti: quindi l’uso del registro RBI come fonte dati permette di garantire *by design* questo aspetto di sicurezza durante tutto l’iter di lavorazione.

Il disegno dell’architettura attuale (versione *as-is*) del processo della base ha fatto emergere la numerosità delle fonti utilizzate, il numero di attori coinvolti, le interconnessioni con le altre basi, rispettivamente base “Popolazione insistente giornaliera per studio e lavoro - Archimede” e la base dati “Precarietà lavorativa - Archimede” e la complessità del processo di costruzione della base stessa. In Figura 4.1 è riportata una schematizzazione della versione attuale di questo processo, con il dettaglio della lista delle undici fonti necessarie all’alimentazione del processo nel suo insieme. Le fonti sono raggruppate per tema prevalente: dati reddituali, dati riferiti al grado d’istruzione dell’individuo e dati anagrafici. È stato analizzato e rappresentato il processo che si occupa di acquisire le fonti amministrative: la maggior parte delle fonti provengono dal Sistema integrato dei microdati, ed alcune di esse vengono dalle altre basi di Archimede, in particolare da “Precarietà lavorativa”. L’intero processo di produzione della base “Condizioni socio-economiche delle famiglie”, è stato scomposto nei seguenti macro processi:

- a. il processo di “Acquisizione dei dati da fonte e da Basi Archimede”, che è possibile scomporre in ulteriori tre sotto-processi che rappresentano le attività di acquisizione e

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

pretrattamento dei dati. I sotto-processi hanno la funzione di rendere i dati armonizzati e integrabili con le altre forniture; i tre sotto-processi riguardano:

- l'acquisizione ed il pretrattamento dei dati di fonte amministrativa da Sistema integrato dei microdati - SIM;
- l'acquisizione dei dati da altri processi Archimede;
- l'acquisizione del Casellario dei pensionati - Inps, nella versione rielaborata e processata dagli esperti tematici.

b. il processo di elaborazione e costruzione della base denominato in Figura 4.1 con nome di “*Determina i dati sulle condizioni socio-economiche delle famiglie*”, è stato scomposto in sette sotto-processi o fasi. Nel dettaglio ritroviamo:

- una [Fase1] di individuazione del dominio della popolazione degli individui di riferimento;
- una [Fase2] di integrazione delle informazioni reddituali e creazione di variabili derivate;
- una [Fase3] di integrazione per ogni individuo delle informazioni su lavoro e titolo di studio e creazione di ulteriori variabili derivate;
- una [Fase4] di costruzione dell'associazione della popolazione di riferimento nelle famiglie e definizione di nuove variabili derivate;
- una [Fase5] di integrazione per ogni famiglia e per ogni suo componente delle variabili derivate dal reddito delle Famiglie;
- una [Fase6] di aggiornamento e integrazione delle variabili Famiglia;
- una [Fase7] di definizione di nuove variabili derivate (associazione del codice regione e del codice ripartizione) e creazione di identificativi statistici (necessari alle successive fasi di diffusione dei dati).

Infine, l'ultimo elemento rappresentato nella modellazione di Figura 4.1 è il prodotto finale, ossia la base dati “Condizioni socio-economiche delle famiglie”, output del macroprocesso.

Dall'analisi della struttura del processo, di tutte le attività svolte e delle fonti di alimentazione necessarie è stato modellato il processo di integrazione della base Famiglie con il SIR.

La proposta relativa alla ristrutturazione del nuovo processo di costruzione della base “Condizioni socio-economiche delle famiglie” (versione *to-be*) è riportata in Figura 4.2: da essa si evince come il processo vada a semplificarsi notevolmente rispetto al processo attuale.

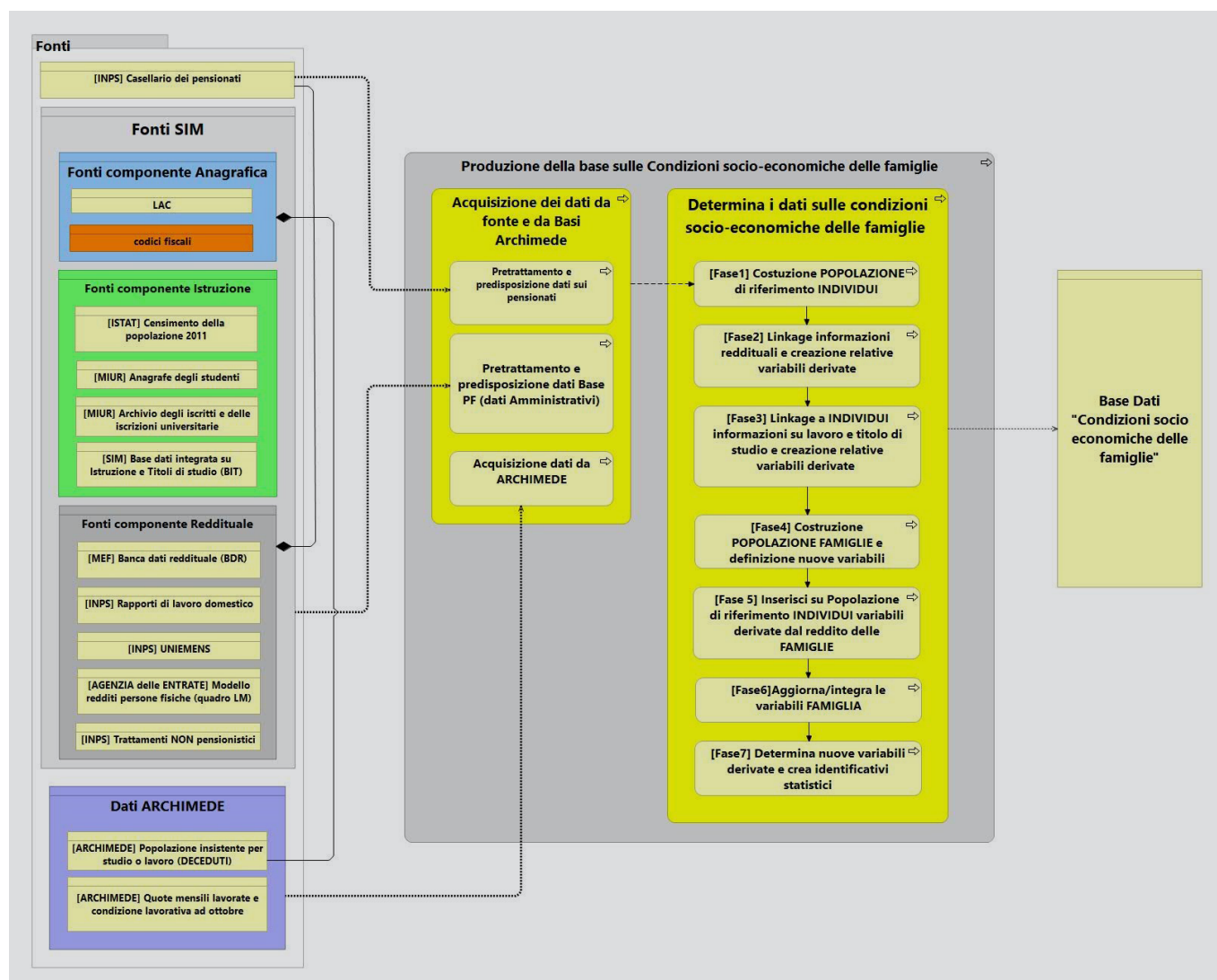
Infatti è possibile osservare come si riducano le fonti necessarie al processo e come l'uso del RBI consenta di:

- a. sostituire integralmente l'uso della fonte Liste anagrafiche comunali (Lac) usata per la costruzione delle popolazioni degli Individui e delle famiglie;
- b. eliminare la necessità di prelevare dalla base “Popolazione insistente giornaliera per studio e lavoro - Archimede” l'informazione sui Decessi, in quanto già inclusa ed armonizzata all'interno del Registro base degli individui. Questo intervento consente di disaccoppiare i due processi di costruzione delle singole basi garantendone una maggiore indipendenza e libertà di azione e di manutenzione su entrambi i processi;
- c. sostituire il processo di definizione del titolo di studio con l'acquisizione di questo da RBI;
- d. eliminare il processo di acquisizione delle seguenti fonti: Base dati integrata su Istruzione e Titoli di studio (BIT) e Censimento della popolazione 2011, utilizzate nel relativo processo di costruzione del titolo di studio.

Nella schematizzazione del nuovo processo è stata mantenuta l'acquisizione di alcune fonti come ad esempio il “Casellario dei pensionati” prevedendone però l'utilizzo nella versione con codici individuali pseudo-anonimizzati. Inoltre sono riportati tra le fonti i due registri di base che sostituiscono buona parte dei dati da fonte amministrativa come già descritto.

L'uso del Registro RSBL, come già spigato nel Paragrafo 4.1, è necessario per prelevare informazioni su province, regioni e ripartizioni territoriali, utilizzate in fase di diffusione dei dati ed inoltre per tenere allineata la codifica del territorio rispetto al tempo.

Figura 4.1 - As-is schema generale del processo relativo alla base "Condizioni socio-economiche delle Famiglie - Archimede"



Fonte: Elaborazione degli autori

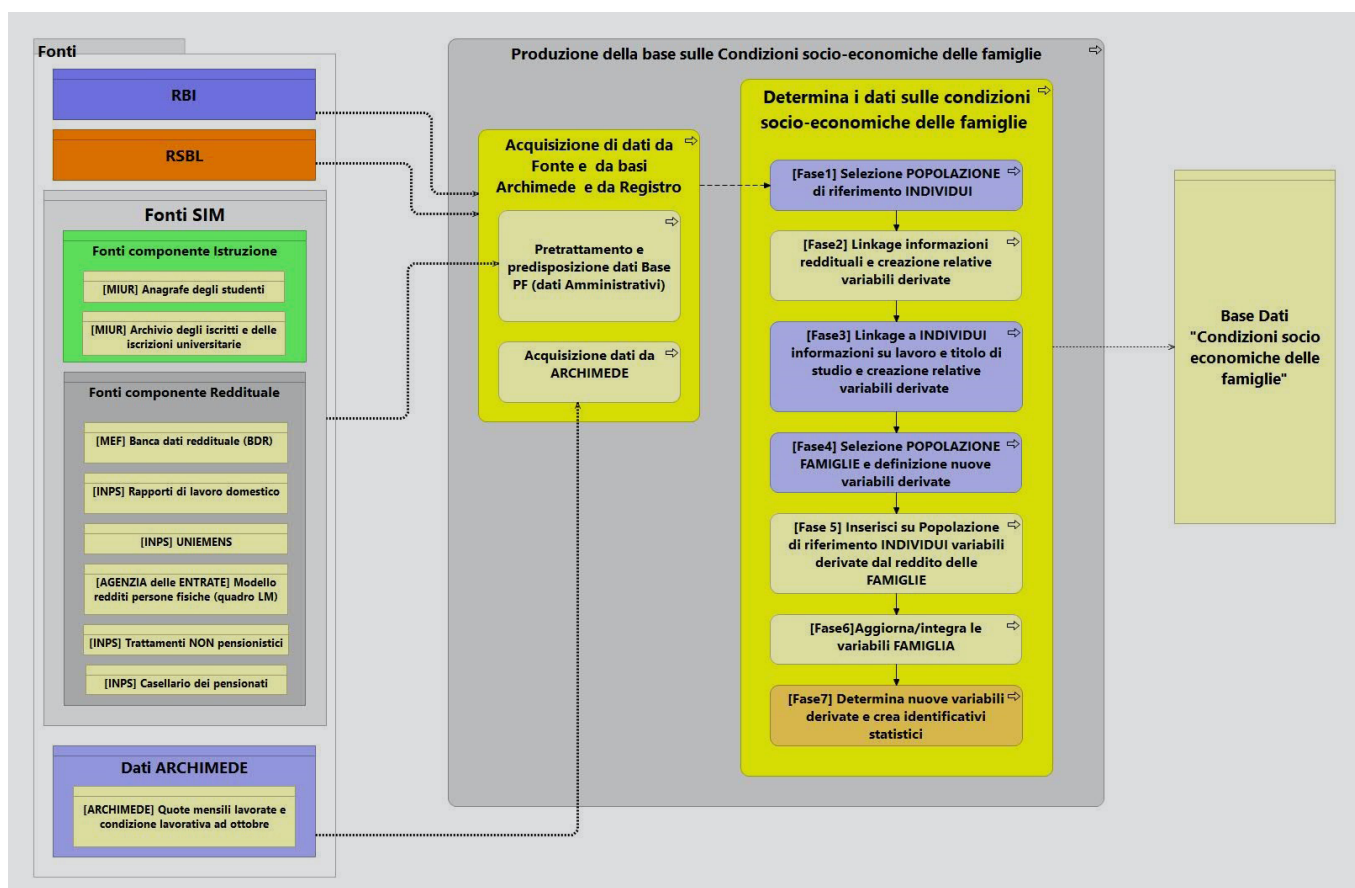
Dopo aver analizzato la possibile semplificazione delle fonti di dati necessarie alla costruzione della base Famiglie riportiamo alcune considerazioni in merito alla semplificazione che otterremmo sul fronte dei processi di costruzione della base Famiglie, infatti:

- dal processo di "Acquisizione dei dati da fonte e da Basi Archimede" viene rimossa la componente afferente all'elaborazione sui dati del Casellario dei pensionati;
- nel processo di elaborazione principale di "Determina i dati sulle condizioni socio-economiche delle famiglie" è ipotizzabile una semplificazione di varie fasi di lavorazione, tra le quali:
 - la [Fase1] in cui non sarà più necessaria l'attuale elaborazione sulle LAC per costruire la popolazione degli individui vivi residenti in famiglia e successivamente

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

l'integrazione della lista dei Deceduti provenienti dalla base "Popolazione insistente giornaliera per studio e lavoro - Archimede", ma sarà sufficiente selezionare dal Registro base degli individui la popolazione di interesse, questo porterà ad

Figura 4.2 - To-be schema generale del processo relativo alla base "Condizioni socio-economiche delle Famiglie - Archimede"



Fonte: Elaborazione degli autori

una nuova versione della [Fase1] descritta nel nuovo modello con il processo: *"Selezione della popolazione di riferimento Individui"*;

- lo stesso tipo di semplificazione è previsto nella [Fase4] dove il processo di costruzione della popolazione residente in famiglia potrà essere semplificata andando ad attingere alle informazioni del registro RBI, che fornisce già la lista della popolazione residente in famiglia e le relative variabili;
- la [Fase3] in cui è prevista l'integrazione delle informazioni su lavoro e titolo di studio per gli individui e la creazione di ulteriori variabili derivate, verrà semplificata per la quota parte relativa all'individuazione del titolo di studio degli individui che come anticipato può essere selezionata dall'RBI evitando di dover integrare i dati provenienti dalla Base integrata dei titoli di studio e dal Censimento della popolazione 2011.

Nella schematizzazione del nuovo processo di Figura 4.2 sono state rappresentate le ulteriori sotto-fasi della nuova architettura. In particolare nella [Fase7] vengono aggiunte le informazioni territoriali dalla fonte Registro RSBL.

4.2.2 Modellazione del processo di produzione della Base Popolazione Insistente giornaliera per studio e lavoro - Archimede armonizzata con il SIR

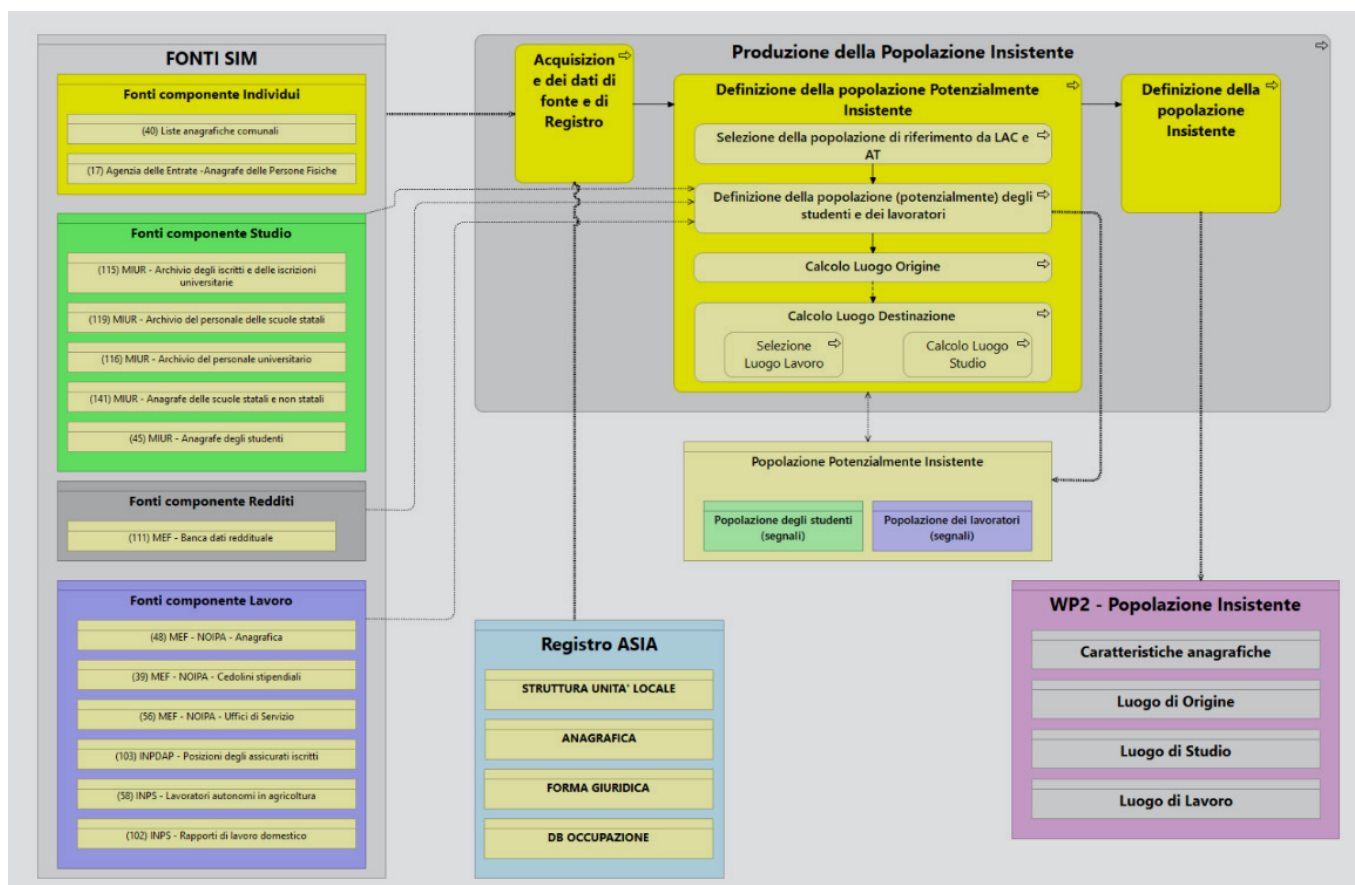
Nella versione attuale il processo di costruzione della base “Base Popolazione Insistente per studio e lavoro - Archimede” descritta nel Capitolo 2, e illustrato in Figura 4.3, è composto di tre fasi: due di caricamento dati ed una di creazione della popolazione obiettivo della base, ossia la popolazione insistente. Tali fasi si possono articolare nelle seguenti macro attività:

- Acquisizione delle fonti amministrative dal Sistema integrato dei microdati;
- Creazione di una prima versione della popolazione insistente (popolazione potenzialmente insistente);
- Caricamento dall'Archivio statistico delle imprese attive (Asia) e dall' Archivio statistico delle unità locali delle informazioni sui lavoratori;
- Determinazione dei comuni di origine e di destinazione degli spostamenti casa/lavoro e costruzione della popolazione insistente.

La popolazione potenzialmente insistente viene creata in due fasi:

- Si seleziona la lista degli individui partendo dalla Lac come fonte primaria, e integrando se necessario i dati anagrafici dalla Banca dati statistica reddituale (Bdr) e dalla Base dati anagrafica dell'anagrafe tributaria sulle persone fisiche (At);

Figura 4.3 - As-is schema generale del processo relativo alla base “Popolazione insistente giornaliera per studio e lavoro - Archimede”



Fonte: Elaborazione degli autori

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

- b. Si analizzano i segnali di presenza nelle fonti relative allo studio ed al lavoro. Se il segnale individuale non è deducibile dall'integrazione, esso viene aggiunto;

Si ottiene la popolazione potenzialmente insistente, che viene corretta per eliminare errori quali individui deceduti e individui duplicati.

La successiva fase di elaborazione della popolazione oggetto delle analisi è composta dai seguenti passi:

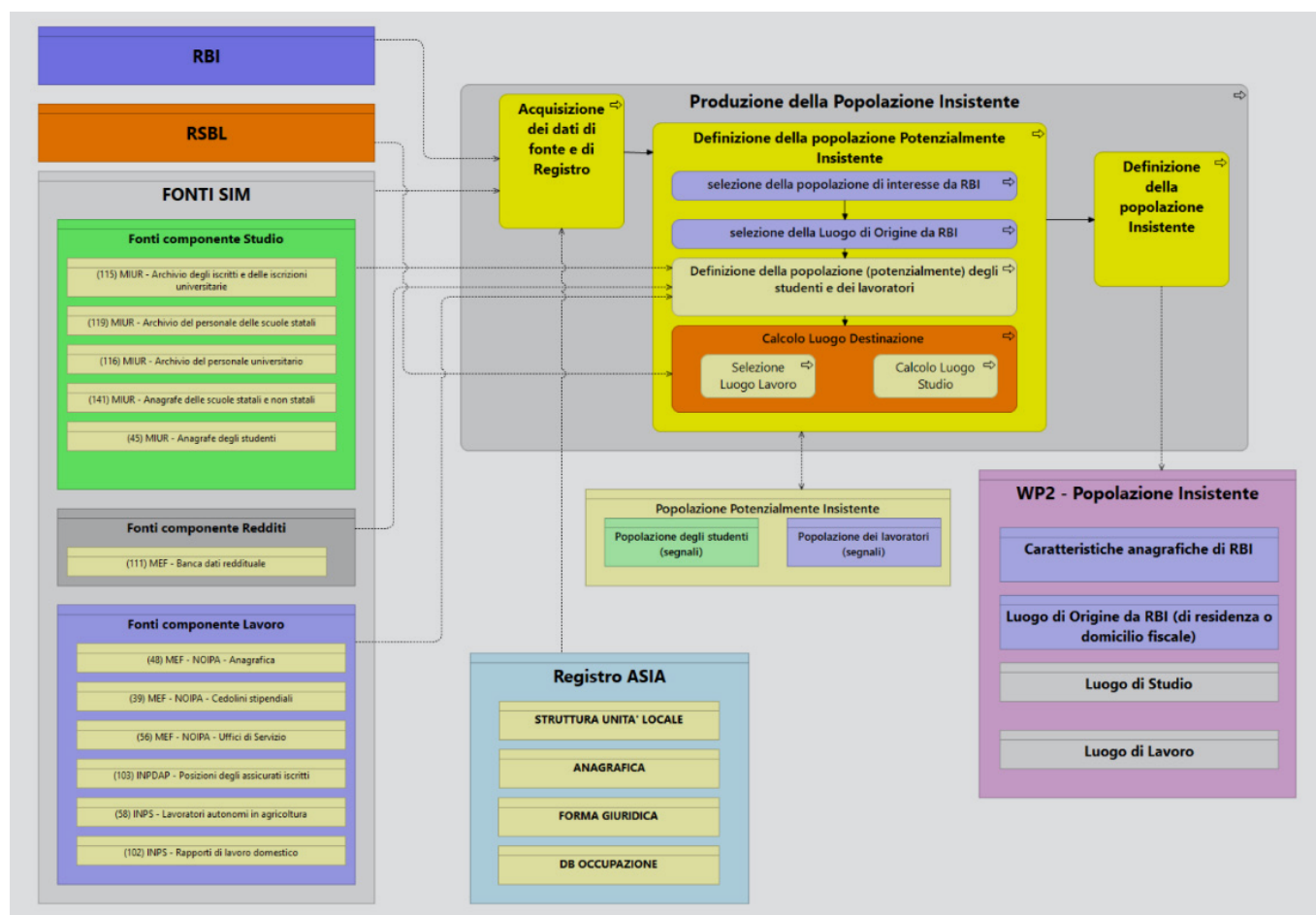
- a. Controllo e deduplicazione degli individui;
- b. Rimozione degli individui deceduti e creazione dell'output della popolazione deceduta come sottoprodotto, successivamente utilizzato nel processo della base Archimede "Condizioni socio-economiche delle famiglie";
- c. Calcolo del Comune di origine;
- d. Calcolo del Comune di destinazione del lavoro.

Nel nuovo modello del processo si prevede di integrare i dati di RBI per la definizione degli individui che entrano nel campo di osservazione della base "Popolazione insistente giornaliera per studio e lavoro - Archimede". In questo caso l'alimentazione da RBI sostituisce completamente l'alimentazione delle fonti anagrafiche e il processo di definizione della popolazione potenzialmente insistente.

In particolare l'uso di RBI elimina tutta la fase di integrazione, controllo, deduplicazione e individuazione dei deceduti, in quanto il dato di RBI ha già operato tutte queste elaborazioni. Inoltre RBI non include soltanto la popolazione residente ma anche la popolazione che pur non essendo residente ha dei segnali di presenza sul territorio nazionale per studio o per lavoro, così come la base di Archimede. Infine il registro è in grado di fornire il Comune di origine sia per i residenti che per i non residenti. Il comune di lavoro invece resta prerogativa della base Popolazione insistente che ha messo a punto un processo specifico per la sua determinazione.

Il trattamento della storicizzazione del territorio operata dal registro del Territorio (RSBL) renderà la codifica dei comuni automatica in funzione del tempo di riferimento dei dati. In Figura 4.4 viene riportato l'aggiornamento del processo e le attività coinvolte.

Figura 4.4 - *To-be* schema generale del processo relativo alla base “Popolazione insistente giornaliera per studio e lavoro - Archimede”



Fonte: Elaborazione degli autori

4.2.3 Modellazione del processo di produzione della Base Precarietà Lavorativa - Archimede armonizzata con il SIR

Il processo di produzione della base "Precarietà Lavorativa - Archimede" nella sua versione attuale è stato descritto nel Paragrafo 3.4; in particolare nei diagrammi sono stati illustrati gli attori, i ruoli, gli input e gli output del processo. In Figura 4.5 è riportata una schematizzazione semplificata che evidenzia l'impatto dell'armonizzazione della base con i dati del SIR in particolare con l'integrazione del Registro base degli individui e delle convivenze e del Registro statistico di base dei luoghi.

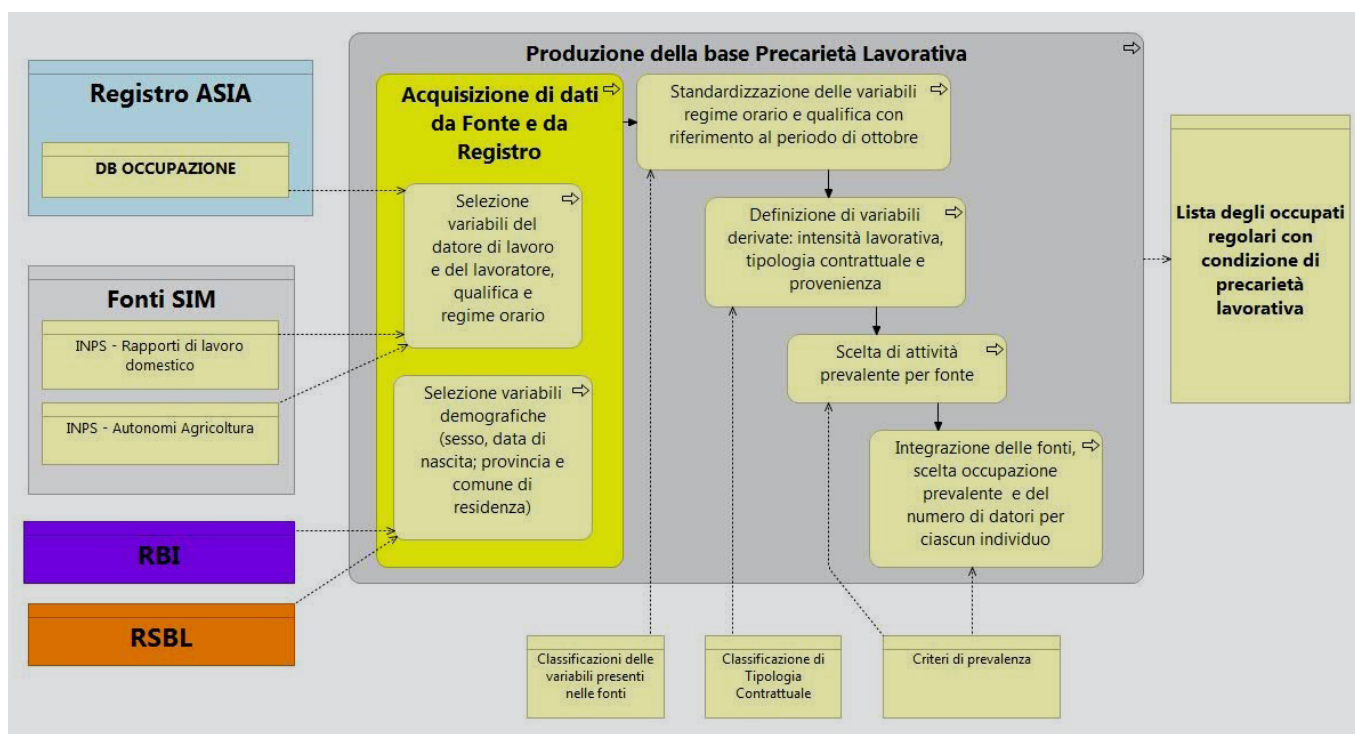
Infatti, partendo da tale schematizzazione, la progettazione del nuovo processo di produzione si è focalizzata nell'evidenziare l'effetto sui processi che vengono revisionati, in particolare quelli dovuti alla sostituzione dell'uso delle Liste anagrafiche comunali per la definizione della popolazione di riferimento. In particolare il processo maggiormente coinvolto dal cambio di alimentazione della base è il sotto-processo di "Selezione delle variabili

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

demografiche” all’interno della macrofase “*Acquisizione di dati da Fonte e da Registro*” così come evidenziato nella Figura 4.5.

Nella versione evolutiva del processo di produzione, per estrarre tali variabili sarà necessario accedere al Registro base degli individui e delle convivenze e contestualmente al Registro statistico di base dei luoghi per l'estrazione della provincia di residenza dei singoli individui selezionati.

Figura 4.5 - *To-be* schema generale della base “Precarietà Lavorativa - Archimede”

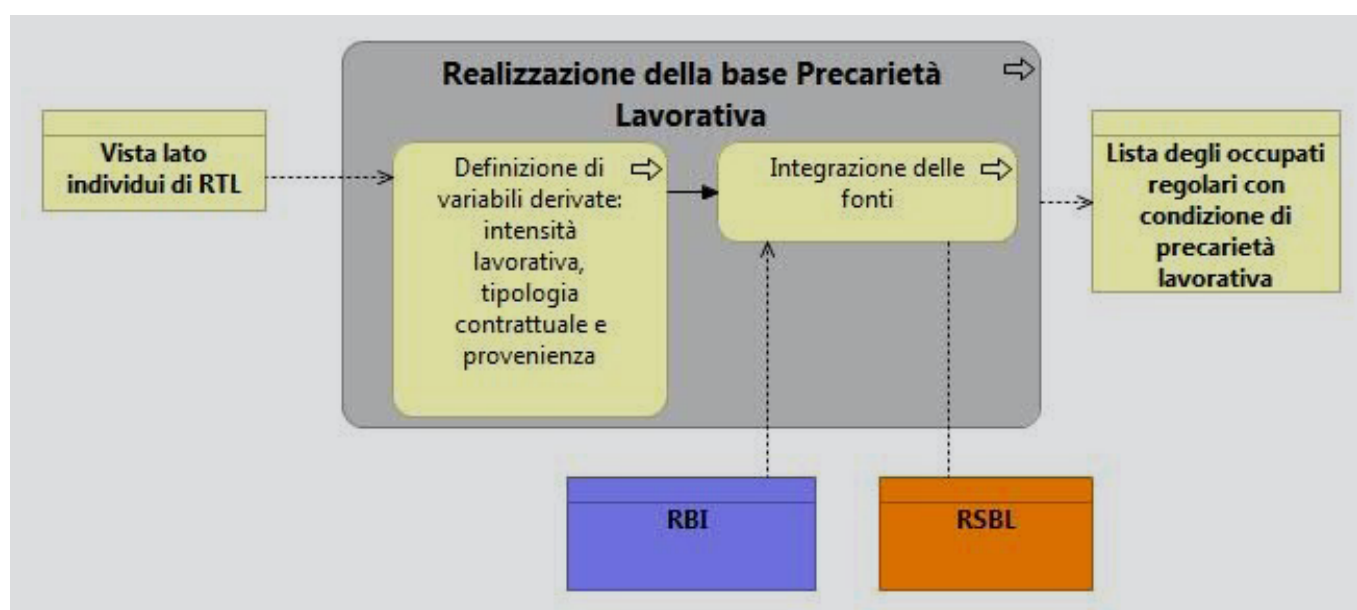


Fonte: Elaborazione degli autori

Per la base “Precarietà lavorativa - Archimede” si può pensare ad un’ulteriore integrazione con un prodotto che sarà fornito dal Registro tematico del Lavoro (RTL) che conterrà tutte le informazioni sugli occupati presenti nelle fonti amministrative con la relativa attività prevalente. Questo prodotto è in fase di completamento e potrà sostituire la fornitura del Registro Asia Occupazione e gran parte del processo di definizione della condizione lavorativa dell’individuo. In questo modo l’intero processo sarà armonizzato con il SIR e non prevederà l’acquisizione di ulteriori fonti amministrative. Inoltre si semplificherà notevolmente riducendosi a due sottoprocessi di selezione delle variabili di interesse e di definizione della situazione lavorativa regolare e la relativa continuità, oltre che le condizioni contrattuali.

La semplificazione dei processi, illustrata in Figura 4.6, sarà possibile in quanto la scelta di attività prevalente per fonte viene eseguita nella produzione di una specifica fornitura di dati del RTL riferita agli individui che svolgono una qualsiasi attività lavorativa. A questo punto resterà a carico della base in oggetto l'integrazione dei dati sull'occupazione degli individui con le informazioni demografiche fornite da RBI e la determinazione delle variabili di interesse della base "Precarietà lavorativa - Archimede".

Figura 4.6 - *To-be* fasi successive schema generale del processo "Precarietà lavorativa - Archimede"



Fonte: Elaborazione degli autori

4.3 Scenari evolutivi dell'armonizzazione di Archimede con il Sistema integrato dei registri (SIR)

Nell'attuale versione di Archimede il processo *Condizioni socio economiche delle famiglie* è dipendente dai processi di produzione delle altre basi poiché utilizza alcuni dei loro output. Per questo motivo i tre processi non sono indipendenti e vanno gestite sia la temporizzazione che le dipendenze, come mostrato in Figura 4.7.

Nello specifico *Condizioni socio economiche delle famiglie* necessita dei seguenti componenti:

- Lista deceduti, prodotta da Popolazione insistente;
- Intensità lavorativa mensile (quota del mese lavorata) e condizioni lavorative dalla base sulla precarietà lavorativa.

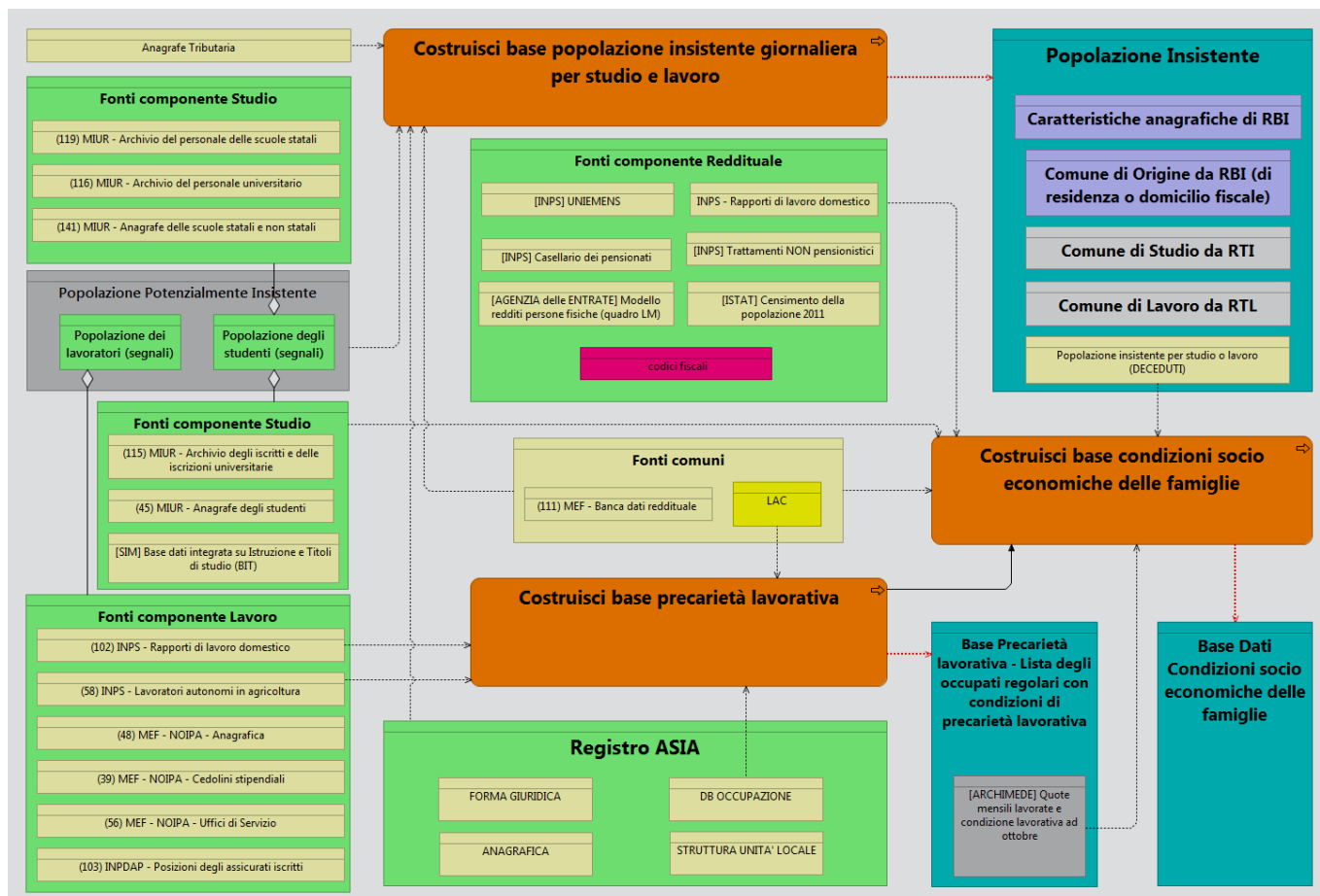
Inoltre l'insieme delle fonti dai quali i tre processi attingono è quasi sempre comune, ma i singoli caricamenti sono specifici per ogni Base. Ciò comporta la definizione di un processo di caricamento, costruzione ed integrazione ad hoc per ogni base. Questo si potrebbe evitare acquisendo i dati già integrati nel SIR, lasciando più margine alle fasi di elaborazione e di determinazioni delle variabili proprie di ciascuna delle basi di Archimede. In questo modo il processo si semplifica e risulta standardizzato, modulare, scalabile e con un controllo di qualità dell'input comune alle tre basi.

La nuova architettura di Archimede lo vede integrato nel quadro generale del SIR come un sistema che crea nuova conoscenza, alimentato dal SIR e che nello stesso tempo si configura come un output predefinito del SIR e che ne costituisce uno dei sistemi di predisposizione di dati per la diffusione dell'istituto.

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

Nella Figura 4.7 viene presentata la vista d'insieme delle tre basi che compongono Archimede così come attualmente è progettato, con le interazioni tra le basi e la loro alimentazione.

Figura 4.7 - Vista generalizzata dei processi di Archimede *as-is*



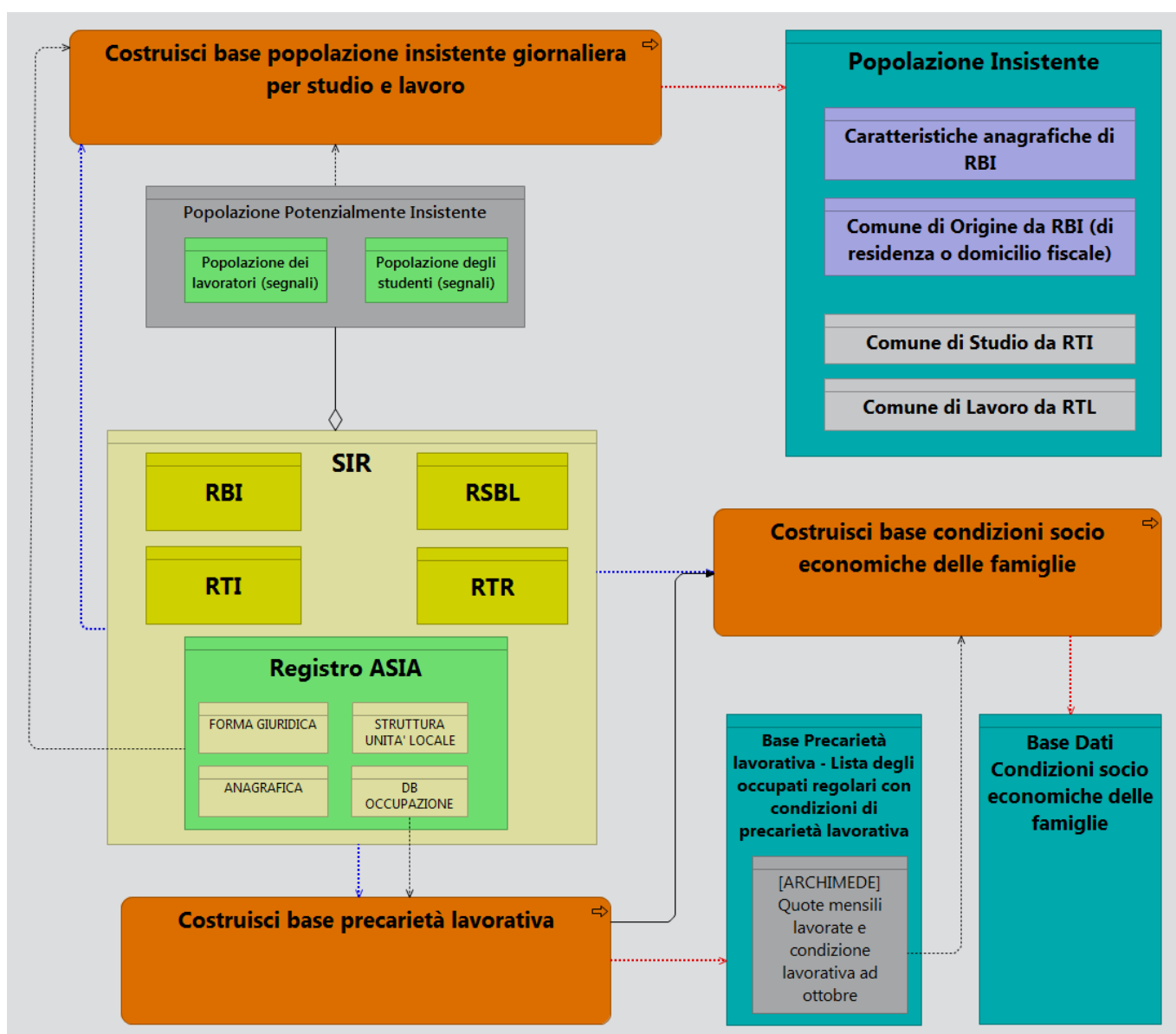
Fonte: Elaborazione degli autori

Nel diagramma di Figura 4.8 è possibile apprezzare come l'integrazione dei processi si semplifichi notevolmente: la nuova modalità di alimentazione consente di derogare in parte o completamente la fase di integrazione dei dati anagrafici e territoriali che possono essere estratti facilmente dai Registri.

Inoltre viene eliminata la dipendenza delle basi per alcuni sotto prodotti della base “Popolazione insistente giornaliera per studio e lavoro - Archimede”, migliorando l'isolamento dei processi e garantendo una maggiore modularità del sistema, una migliore resa qualitativa del dato, già validato dal registro competente.

Quando il processo di costruzione del SIR sarà completato con l'inclusione del registro tematico dell'istruzione (RTI) e dei Redditi (RTR) e con il completamento del registro tematico del Lavoro (RTL) e della sua componente sugli individui occupati come descritto nel Paragrafo 4.2.3, si potrà avere una configurazione come illustrato nella Figura 4.8.

Figura 4.8 - Vista generalizzata dell'integrazione dei processi di Archimede con SIR



Fonte: Elaborazione degli autori

La nuova configurazione del processo Archimede sarà completamente sgravato dalla funzione di acquisizione e di integrazione delle fonti. Questo sarà comunque possibile soltanto a seguito di analisi di qualità degli altri registri del Sistema integrato dei registri,

4. Proposta di una nuova architettura armonizzata con il Sistema integrato dei registri (SIR) delle basi di Archimede

e dopo una attenta valutazione della sostituibilità in funzione dei miglioramenti che si otterrebbero modificando l'alimentazione dei processi di costruzione delle singole basi.

Un discorso particolare si può fare per la base "Precarietà lavorativa - Archimede", per la quale si può pensare una ulteriore integrazione con un prodotto fornito dal Registro del Lavoro che corrisponde agli individui occupati. Questo prodotto al momento è utilizzato per il processo di definizione dell'occupazione del Censimento e consiste in una vista lato individui delle posizioni lavorative dei lavoratori che fornisca tutte le variabili necessarie al processo di produzione della base, ovvero le informazioni sugli occupati presenti nelle fonti amministrative con la relativa attività prevalente.

Questo prodotto in fase di completamento all'interno di RTL potrà sostituire la fornitura del Registro Asia Occupazione e gran parte del processo di definizione della condizione lavorativa dell'individuo. Se si deciderà di armonizzare la base anche con questa fornitura di RTL, allora l'intero processo della base sarà armonizzato con il SIR e non prevederà l'acquisizione di alcuna fonte amministrative. Quindi, come già introdotto nel paragrafo 4.2.3, si semplificherà notevolmente riducendosi a due sottoprocessi di selezione delle variabili di interesse e definizione delle stime inerenti la situazione lavorativa regolare e la relativa continuità, e le relative condizioni contrattuali.

4.4 Valutazione di impatto della nuova architettura e proposta di aggiornamento in fasi

La nuova progettazione del sistema di basi Archimede ha evidenziato come l'integrazione completa del processo di costruzione delle basi con il SIR renda il processo meno ridondante, standardizzato e soprattutto sollevato dal processo di acquisizione di una parte dei dati grezzi. L'armonizzazione di Archimede con il SIR renderebbe le basi allineate anche dal punto di vista dei metadati e della qualità al sistema dei registri. Inoltre sarebbe arricchito di una architettura di gestione dell'informazione simile a quello del SIR rendendolo un prodotto del Sistema integrato stesso. Questa versione completa dell'integrazione di Archimede con il SIR per alcune informazioni deve attendere il completamento di alcune componenti del sistema di registri. Quindi il processo di integrazione procederà con una successione di fasi che provvederanno all'integrazione delle componenti in funzione del loro stato di completamento e comunque dopo uno studio di sostituibilità delle informazioni e di misurazione del livello di qualità delle stime effettuato dagli esperti delle basi di Archimede.

Come descritto in questo documento, in questa prima fase il lavoro di valutazione della sostituibilità si è concentrata sui dati del RBI e del RSBL perché attualmente sono i primi registri messi in produzione dal SIR, ma entro il 2022 anche il Registro del lavoro sarà consolidato in tutte le sue componenti e con esso la parte della base degli occupati che contiene gli individui lavoratori corredati di tutte le caratteristiche occupazionali e contrattuali, oltre che al luogo di lavoro.

In generale la reingegnerizzazione delle basi Archimede prevede un processo in fasi che differiscono per impatto nell'aggiornamento del software di integrazione e naturalmente nei tempi di realizzazione, e di stabilizzazione dell'integrazione con l'intero SIR.

La prima fase prevede l'acquisizione del registro RBI come alimentazione delle basi Archimede in sostituzione delle fonti anagrafiche. La realizzazione del cambio di alimentazione comporta la costruzione di un componente di raccordo tra la vecchia alimentazione e la nuova da registro. Questa soluzione provvisoria garantisce:



- a. un basso impatto nella reingegnerizzazione;
- b. la conservazione del software esistente con l'esclusione di alcune componenti come quella di pulizia dei deceduti presenti in LAC;
- c. tempi brevi di realizzazione e godimento dei benefici, se pur parziali, da subito.

Il tutto implica che si rimandi ad una fase successiva la reingegnerizzazione dell'intero processo di Archimede per avere l'integrazione di RBI e anche degli altri registri del SIR.

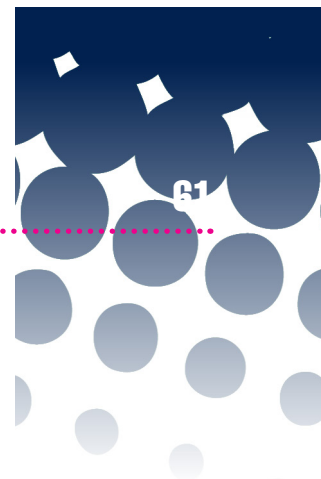
Mentre la seconda fase del progetto prevede la modularizzazione dell'alimentazione delle basi Archimede, che oltre a prevedere in prima istanza l'integrazione di RBI in sostituzione delle fonti anagrafiche, predispone le connessioni per l'alimentazione da altri registri. Questo rende il processo pronto ad acquisire come fonti i registri, che al momento non sono completi, come ad esempio: il Registro del lavoro e la base degli occupati, il Registro dell'istruzione e il Registro dei Redditi, non appena questi saranno rilasciati in produzione. Questa seconda fase di realizzazione dell'architettura del processo Archimede consente di predisporlo all'integrazione con tutti i registri del SIR e lascia la decisione di effettuare l'integrazione soltanto alla valutazione della qualità dei registri del SIR e ai tempi di messa in produzione di questi.

È stata valutata anche l'opzione di realizzare direttamente questa seconda fase, consapevoli che i vantaggi comporterebbero:

- a. una predisposizione unica dell'integrazione delle basi Archimede con il SIR che potrà essere attuato in funzione dei rilasci dei vari registri del SIR, in funzione della valutazione qualitativa dei dati da registro svolta dagli esperti di Archimede.
- b. la ristrutturazione da subito dell'intero ambiente di elaborazione dei dati delle tre basi suddividendole in area di *working* per l'integrazione delle fonti (registri e fonti amministrative) e costruzione dell'output delle singole basi Archimede. La separazione delle due aree è una misura per il miglioramento della gestione dell'impatto sulla Privacy e per la gestione della conservazione dei dati.

L'adozione direttamente di questa seconda architettura che sarà quella definitiva, presenta tempi di realizzazione più lunghi poiché comporta impatti rilevanti sull'implementazione completa di gran parte del software di gestione del processo Archimede ed eventuali piccoli aggiustamenti per integrare registri per i quali ad oggi non è ancora chiara l'architettura realizzativa poiché sono ancora in una fase prototipale.

In funzione di tutte le considerazioni espresse in questo documento, si ritiene preferibile procedere alla realizzazione della prima fase di integrazione con RBI e RSBL e di avviare lo studio di qualità di RTL, monitorando gli stati di avanzamento di altri registri per valutarne l'integrabilità. Inoltre viene avviata la progettazione della creazione dell'infrastruttura completa di integrazione di Archimede con il SIR.



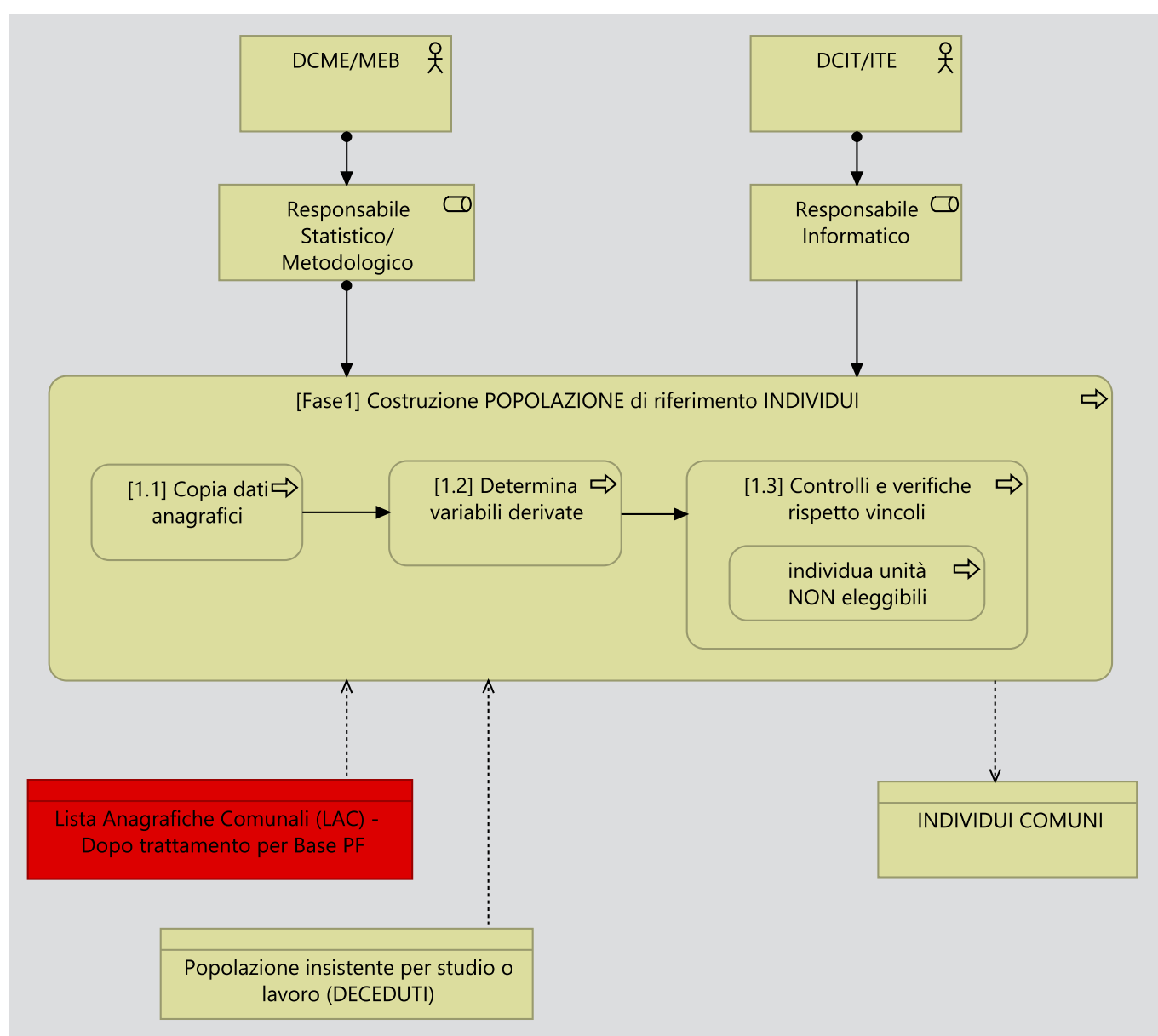
RIFERIMENTI BIBLIOGRAFICI

- Ambroselli, S. 2015. “I codici identificativi univoci all’interno del SIM (Sistema Integrato di Microdati)”. *Istat working papers*, N. 5/2015. Roma: Istat. <https://www.istat.it/it/archivio/156101>.
- Di Zio, M., P.D. Falorsi, e M. Fortini. 2016. “Un sistema di registri statistici basato sull’integrazione di informazioni da archivi amministrativi, indagini statistiche e stime di popolazione”. Presentazione alla Giornata di Studio *MILeS2016*. Milano - *Impresa, Lavoro e Società 2016*.
- Donatiello, G. (a cura di) 2011. “La metodologia di stima dei redditi lordi nell’indagine Eu-Silc. Indagine europea sui redditi e le condizioni di vita delle famiglie”. *Metodi e Norme*, N. 49/2011. Roma: Istat. <https://ebiblio.istat.it/digibib/Metodi%20e%20norme/IST0001667Ed2011N49.pdf>.
- Garofalo, G. 2014. “Il Progetto ARCHIMEDE obiettivi e risultati sperimentali”. *Istat working papers*, N. 9/2014. Roma: Istat. <https://www.istat.it/it/archivio/140232>.
- Istituto Nazionale di Statistica - Istat. 2020. “Popolazione insistente per studio e lavoro. Un approccio sperimentale con dati amministrativi. Anno 2016”. *Statistiche Sperimentali, Analisi e Quadri interpretativi*. Roma: Istat. <https://www.istat.it/it/archivio/239358>.
- Istituto Nazionale di Statistica - Istat. 2018. “A misura di Comune, Sistema informativo statistico multi-fonte”. *Statistiche Sperimentali, Nuovi Indicatori*. Roma: Istat. <https://www.istat.it/it/archivio/220004>.
- Lankhorst, M. 2018. *Enterprise Architecture at Work. Modelling, Communication and Analysis. Fourth Edition*. Cham, Switzerland: Springer Nature, *The Enterprise Engineering Series*.

APPENDICE

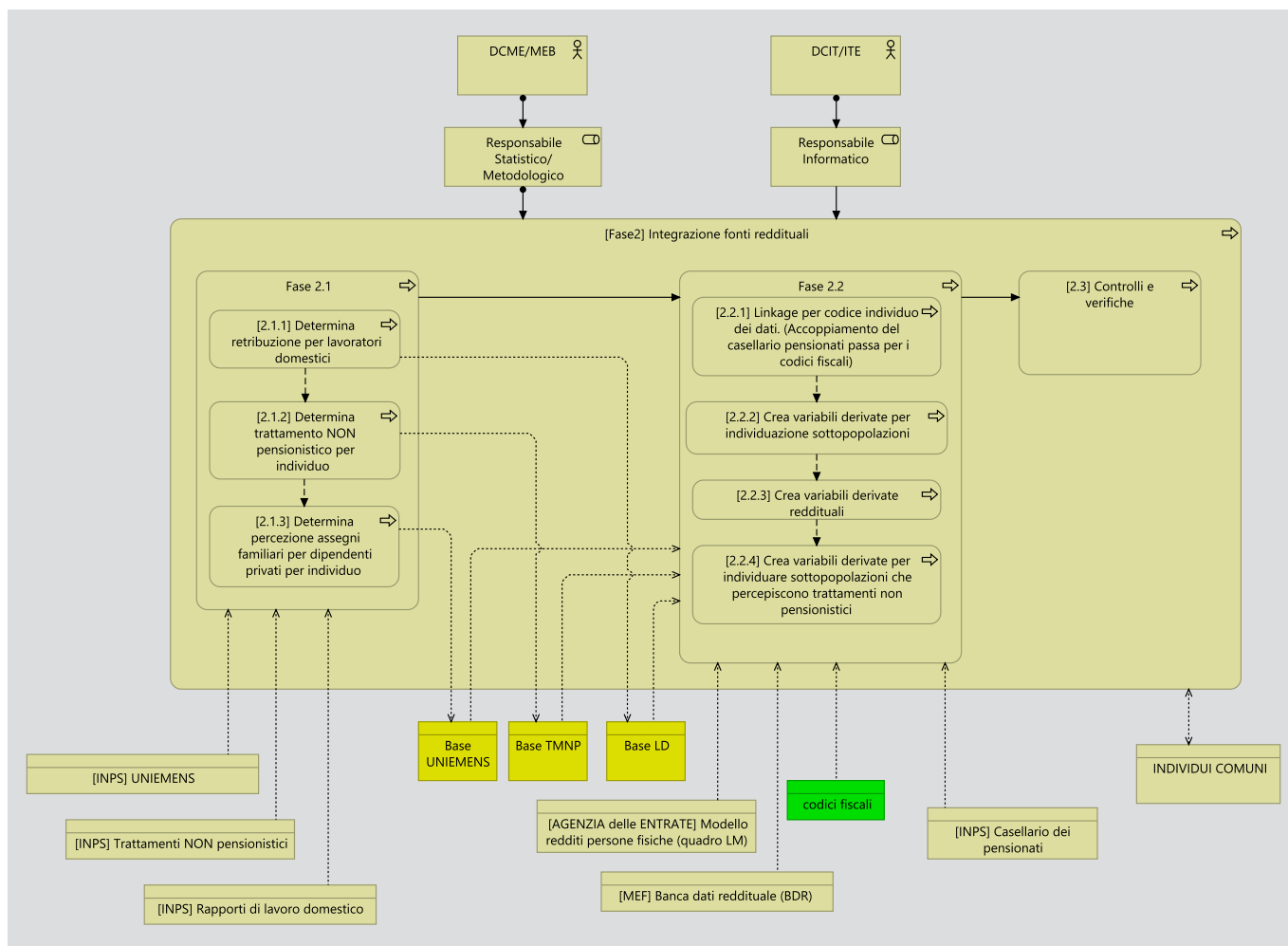
Schemi di dettaglio delle fasi relative alla procedura “Determina i dati sulle condizioni socio-economiche delle famiglie”:

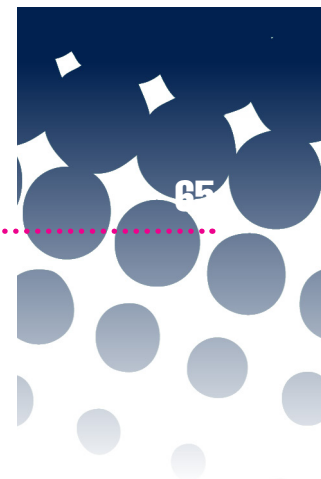
Schema di dettaglio Fase 1



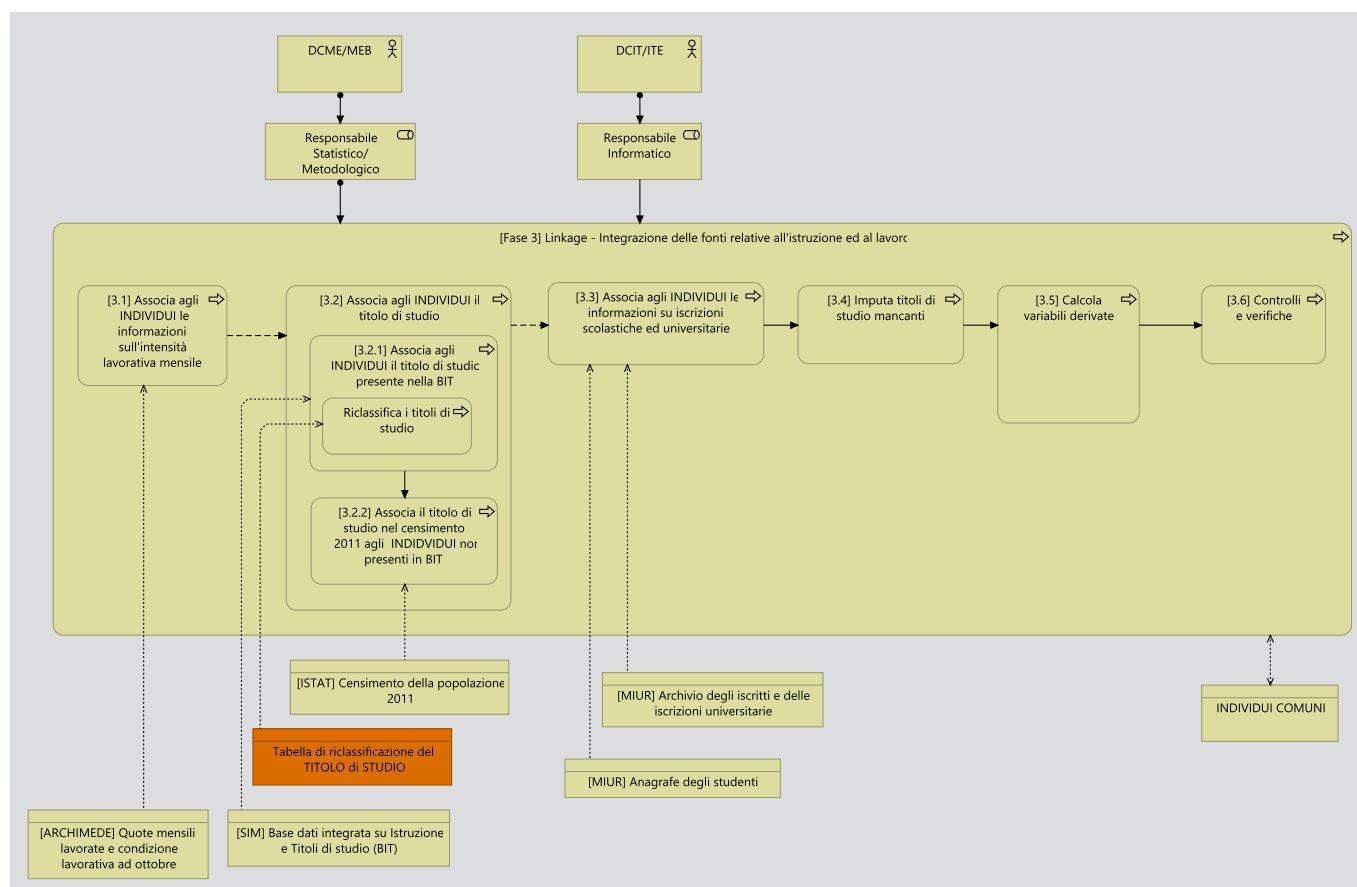
Fonte: Elaborazione degli autori

Schema di dettaglio Fase 2



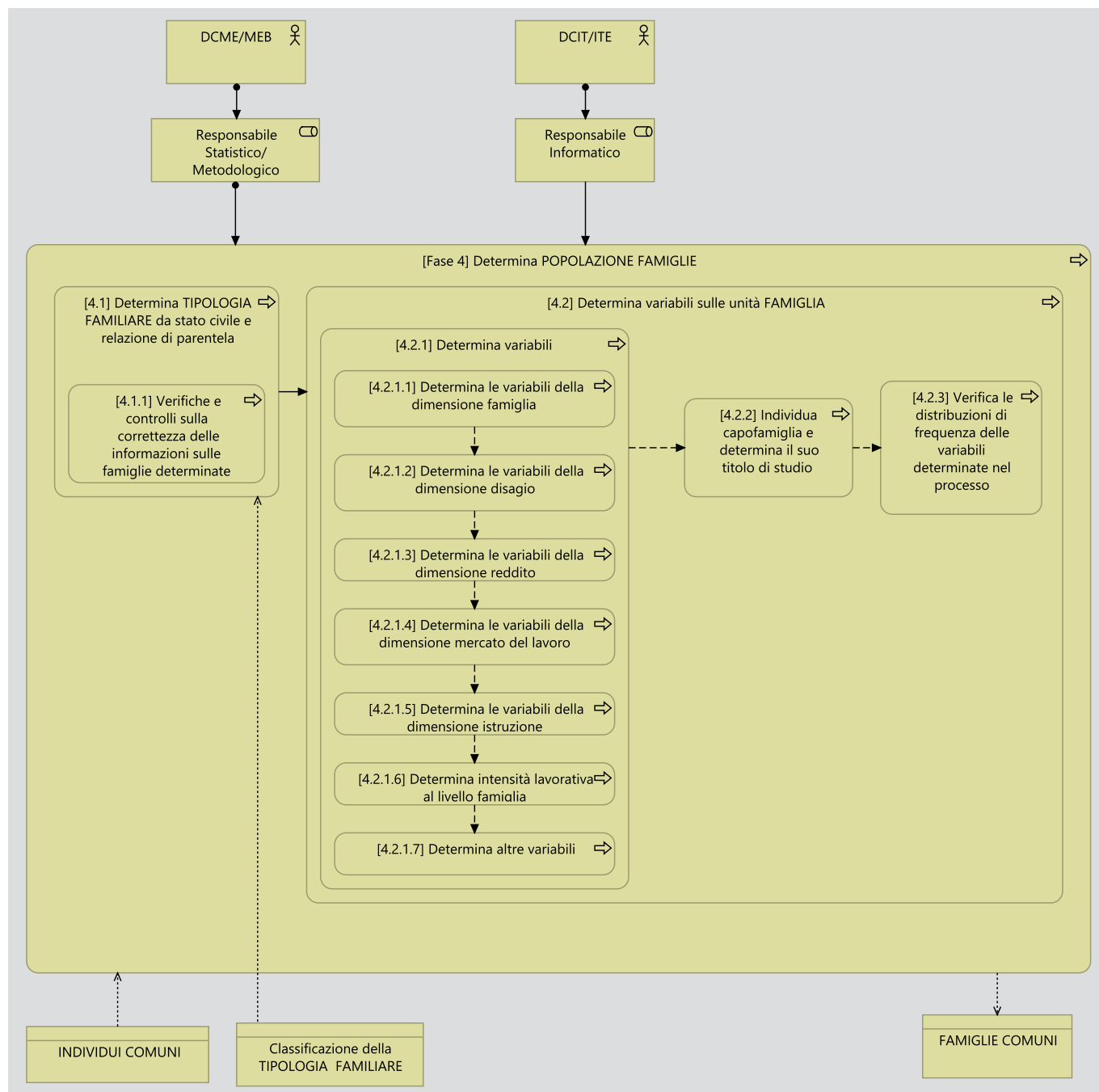


Schema di dettaglio Fase 3



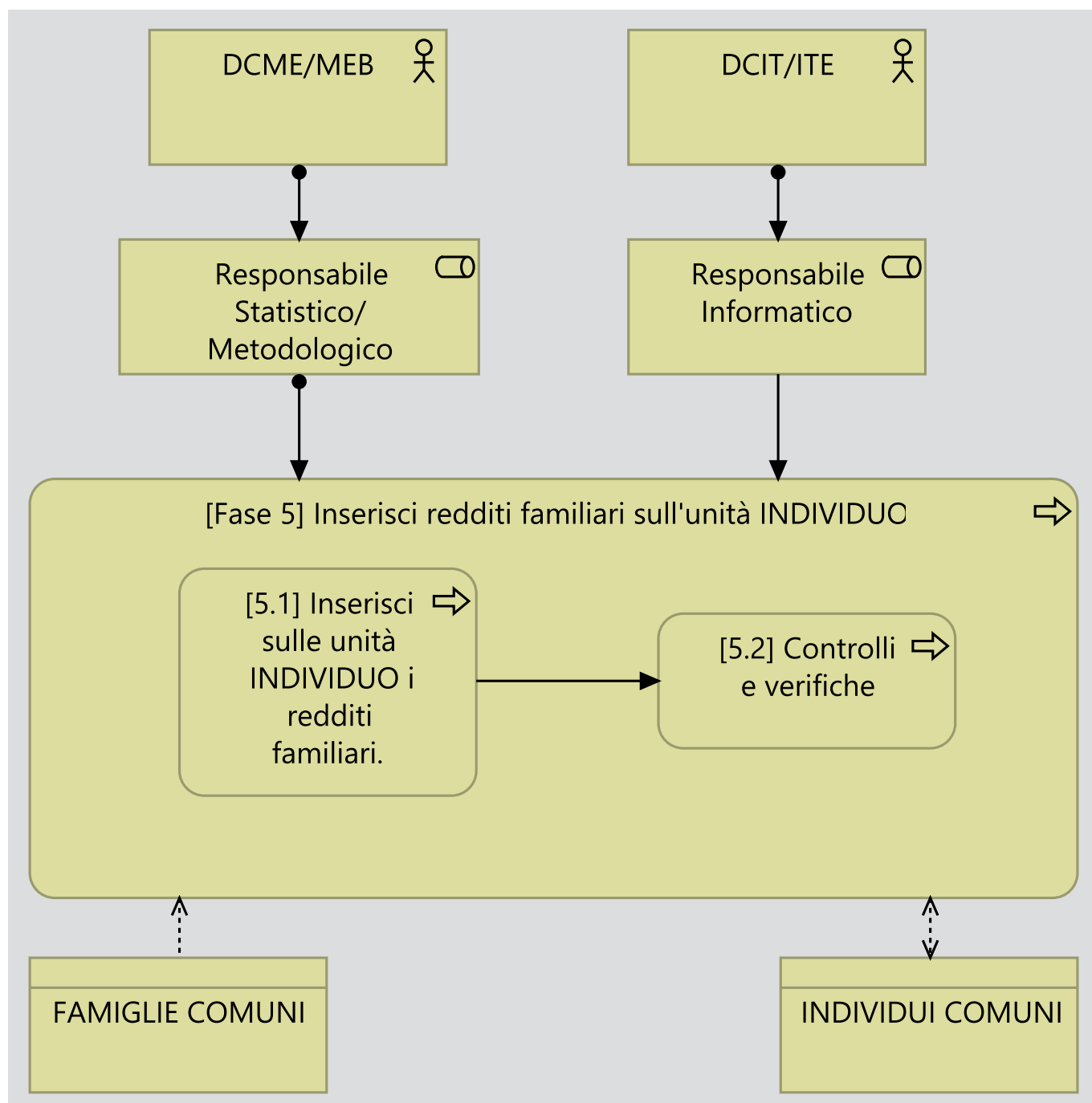
Fonte: Elaborazione degli autori

Schema di dettaglio Fase 4



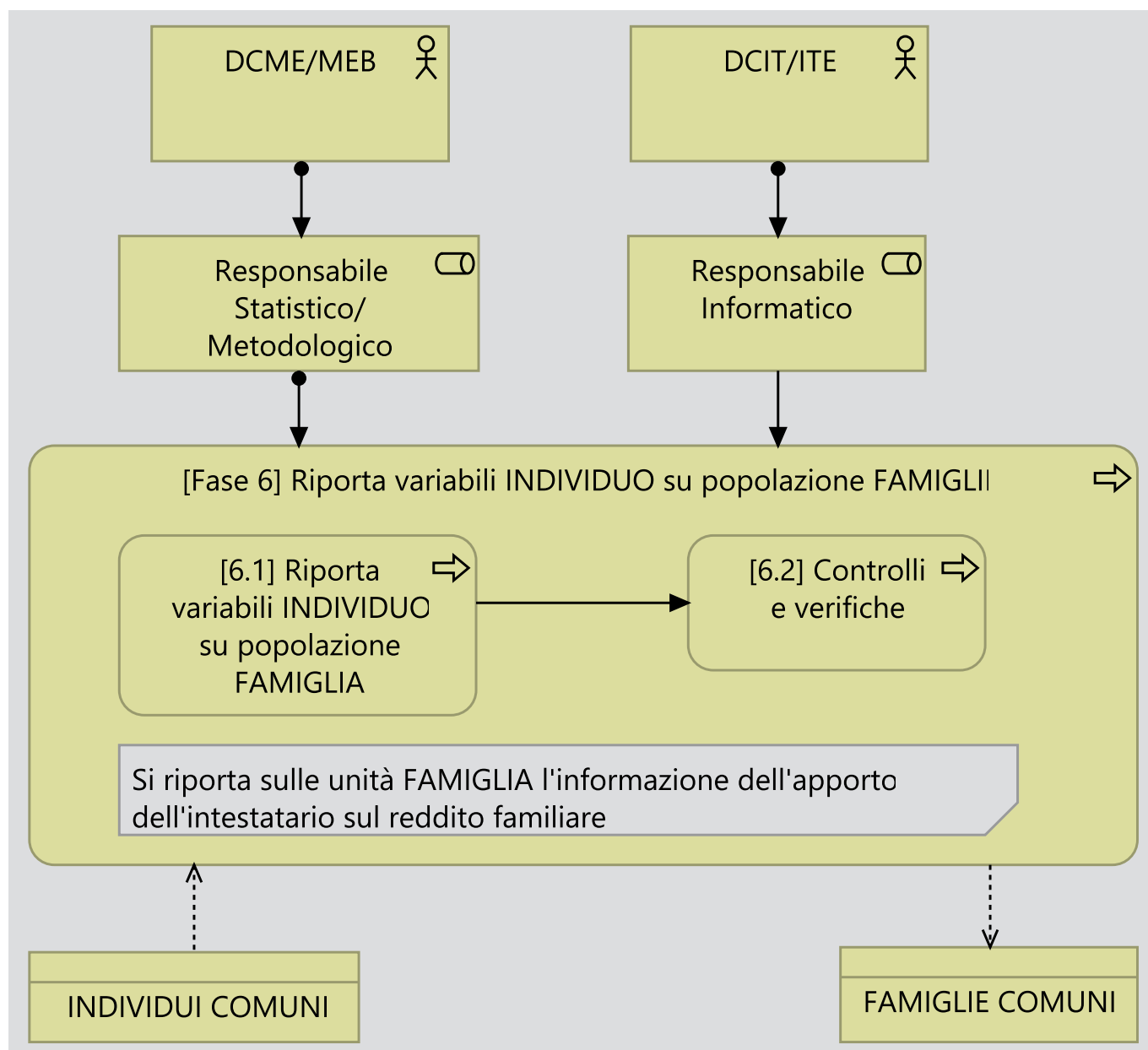
Fonte: Elaborazione degli autori

Schema di dettaglio Fase 5

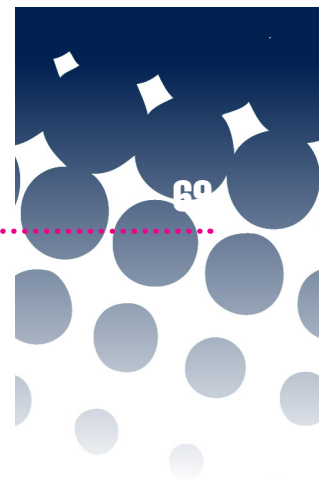


Fonte: Elaborazione degli autori

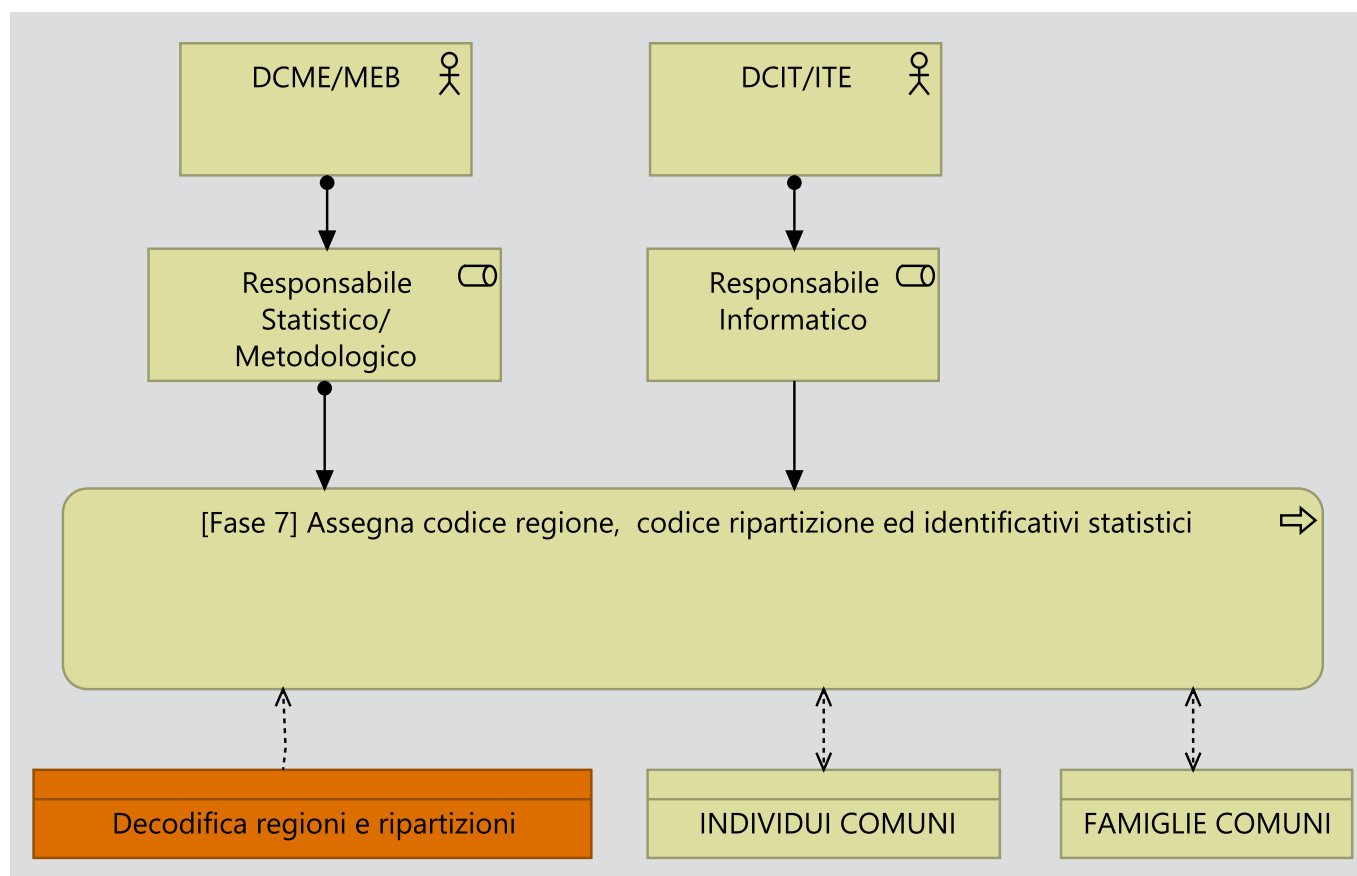
Schema di dettaglio Fase 6



Fonte: Elaborazione degli autori



Schema di dettaglio Fase 7



Fonte: Elaborazione degli autori

