

Editoriale

L'utilizzo delle banche dati amministrative in Italia per informare le politiche

Fabio Berton

European Commission, Joint Research Centre

Roberto Leombruni

Università di Torino

1. Perché i dati amministrativi?

Il fascicolo 2/2026 della rivista *Sinappsi* è il primo di due numeri speciali dedicati allo stato dell'arte sull'utilizzo delle banche dati amministrative a scopo di ricerca in Italia. L'obiettivo è quello di evidenziare ed esemplificare le caratteristiche che li rendono particolarmente adatti all'analisi statistica e alla valutazione delle politiche pubbliche, come testimonia non solo la crescente domanda di accesso a queste banche dati da parte della comunità scientifica, ma anche la crescente attenzione politica a livello italiano ed europeo.

Per dati amministrativi intendiamo *“data collected by public entities for their service activities”* (Crato e Paruolo 2019, v); più precisamente, intendiamo i microdati raccolti da tali enti nell'espletamento delle loro funzioni, che riguardano quindi informazioni a livello del singolo individuo, famiglia, impresa ecc., senza aggregazioni. La “lunga e tortuosa” storia del loro utilizzo a scopo di ricerca scientifica in Italia (Leombruni e Ardito 2026), inizia a metà degli anni Ottanta del secolo scorso quando Bruno Contini, oggi Professore Emerito di Econometria dell'Università di Torino, e Riccardo Revelli – troppo prematuramente scomparso – ebbero l'idea di trasformare le serie dei contributi previdenziali conservate presso l'Istituto nazionale di previdenza sociale (Inps) in storie lavorative individuali, quelle che oggi siamo abituati a chiamare LEED (*Linked Employer-Employee Data*). Da allora questa storia si è evoluta molto, sia in ampiezza – perché oggi gli ambiti di studio coperti dall'informazione amministrativa sono cresciuti ben al di là di quanto contenuto negli archivi dell'Inps – sia in profondità, con nuove forme di accesso ai dati che cercano, con diversi gradi di successo, di coniugare flessibilità di utilizzo per i ricercatori con tutela della riservatezza del dato. Questo numero speciale della rivista *Sinappsi* si propone di dare conto, nel modo più esaustivo possibile, dell'attuale stato dell'arte in Italia dell'utilizzo delle fonti amministrative a scopo di analisi statistica.

Almeno due elementi spiegano la centralità di questo tema nel dibattito corrente. Da un lato, alcune caratteristiche dei microdati di fonte amministrativa li rendono particolarmente adatti all'analisi statistica: (i) poiché vengono raccolti dalle amministrazioni pubbliche nello svolgimento delle proprie attività, i dati amministrativi spesso coprono l'intera popolazione di riferimento, anziché un campione; ciò consente di analizzare con maggiore precisione gli effetti distributivi delle politiche in esame, e quindi di comprendere non solo se una politica funziona, ma anche per chi funziona. (ii) A differenza delle indagini campionarie, i dati amministrativi non dipendono dalla capacità degli individui di ricordare eventi passati né dalla loro disponibilità a riferire informazioni potenzialmente sensibili. Le informazioni elementari vengono infatti raccolte nel corso dell'attività amministrativa, riducendo così alcuni dei tipici problemi di errore di misura, anche involontario. (iii) Anche sotto il profilo economico il loro impiego è vantaggioso: i costi di raccolta sono già sostenuti per la finalità amministrativa originaria. (iv) Inoltre, la presenza di identificativi univoci permette di collegare archivi differenti e, quando possibile, di integrarli con dati provenienti da indagini di tipo survey, ampliando notevolmente la quantità di informazioni disponibili per l'analisi. (v) Infine, la raccolta continua rende questi dati disponibili pressoché in tempo reale e consente di seguire individui, famiglie o imprese per periodi molto lunghi.

Dall'altro lato – e almeno in parte come conseguenza di quanto appena descritto – esiste una crescente domanda, da parte della comunità scientifica e in alcuni casi anche delle stesse pubbliche amministrazioni, di accesso alle fonti amministrative integrate per incrementare la qualità e la quantità della ricerca. De Paola *et al.* nel loro contributo a questo numero monografico ci raccontano che il solo programma VistINPS, operativo dal 2016, ha sostenuto finora oltre 260 progetti di ricerca e la pubblicazione di circa 100 articoli nella collana dedicata. Volgendo lo sguardo fuori dall'Italia, Lima *et al.* (2026) danno conto di un analogo numero di progetti attivati attraverso il Microdata Research Laboratory della Banca del Portogallo (BPLIM), per un totale di circa 350 ricercatori e ricercatrici coinvolti. Il *Centre d'accès sécurisé aux données* (CASD) dell'Istituto nazionale di statistica in Francia ha, dal 2012 in avanti, ospitato circa 800 progetti scientifici per oltre 1.600 ricercatori attivi in tutto il mondo. Ancora, al 2024 le 7.500 basi dati disponibili presso il *Centraal Bureau voor de Statistiek* (CBS) dei Paesi Bassi avevano attratto l'interesse di circa 1.900 ricercatori e ricercatrici (Bloem *et al.* 2026).

Il risultato di questo processo combinato è una crescente attenzione per i dati amministrativi non soltanto da parte della comunità scientifica e degli enti e istituti proprietari del dato elementare, ma anche da parte del mondo politico: sono del giugno 2024, per esempio, le conclusioni del Consiglio dell'Unione europea che invitano la Commissione europea a “*enhance access to administrative data for policy evaluation and impact assessment purposes*” (Consiglio dell'Unione europea 2024, punto 13). La stessa Commissione europea, nel luglio 2026, riconosce che “*Administrative data held by public authorities are a unique source of evidence for policy making and monitoring*” e che “*The use of administrative data can be more efficient and less costly than existing alternatives*” (Commissione europea 2026). Non ci siamo pertanto sorpresi, ma ne siamo stati ugualmente molto compiaciuti, che la *call for papers* per questo numero speciale di *Sinappsi* sull'utilizzo e sull'accesso ai dati amministrativi in Italia abbia raccolto un così ampio riscontro, al punto da richiedere di spezzare la pubblicazione dei testi accettati per la pubblicazione su due fascicoli consecutivi. Insomma, parafrasando un volume edito dallo stesso Bruno Contini assieme a Ugo Trivellato (2005), “Eppur si muove”.

2. Il contenuto del numero monografico

Il fascicolo inizia con un inquadramento giuridico del tema. Il contributo di **di Somma** (*I dati amministrativi: le regole del GDPR per l'uso a fini statistici e di ricerca*) analizza l'assetto normativo italiano ed europeo che disciplina il trattamento dei dati amministrativi a fini statistici e di ricerca, con particolare riferimento al Regolamento generale sulla protezione dei dati (GDPR, Regolamento UE 2016/679). L'autrice parte dalla constatazione che il patrimonio informativo custodito dalle pubbliche amministrazioni è divenuto strategico per le politiche basate sull'evidenza, ma che gran parte delle banche dati contiene informazioni personali, in alcuni casi sensibili. In tale contesto, la normativa non impedisce il loro riutilizzo, ma ne regola rigorosamente le condizioni di liceità, le garanzie da adottare e i limiti di utilizzo. Il testo ricostruisce le origini del trattamento statistico nel Sistema statistico nazionale (Sistan) e delle norme che hanno istituito il Programma statistico nazionale e imposto l'interconnessione dei sistemi informativi pubblici. Vengono delineate le disposizioni che garantiscono il segreto statistico e il divieto di “ricadute amministrative”, vietando l'impiego dei dati raccolti per scopi diversi da quelli statistici. Il Codice della privacy si integra con le regole deontologiche per i trattamenti statistici e le modalità di accesso ai dati da parte dei ricercatori esterni, prevedendo l'obbligo di anonimizzazione o pseudonimizzazione dei dataset. La disciplina è infine integrata dalle successive modifiche che allineano il quadro nazionale al GDPR, introducendo deroghe specifiche per le attività statistiche. Il GDPR, infatti, prevede deroghe specifiche ai principi di limitazione della finalità e di conservazione, nonché esenzioni dal divieto di trattare categorie particolari di dati, se l'utilizzo degli stessi ha finalità statistiche e a condizione che siano adottate misure adeguate di protezione, tra cui la pseudonimizzazione. In conclusione, l'autrice sostiene che, accanto al robusto apparato normativo, l'efficace utilizzo dei dati amministrativi richiede investimenti continui in misure tecniche (pseudonimizzazione, anonimizzazione, *privacyenhancing technologies*) e nel monitoraggio delle pratiche di accesso. Solo un bilanciamento attento tra tutela dei diritti fondamentali e necessità di ricerca, argomenta di Somma, potrà garantire che questi dati continuino a rappresentare un bene pubblico capace di migliorare l'evidence-based policy.

Definito il quadro giuridico, si prosegue con la testimonianza di una delle voci storicamente più importanti sull'uso delle fonti amministrative a scopo statistico, quella di Bruno **Anastasia**. Il suo contributo (*Note su avventure e sventure dei dati amministrativi, con riferimento all'esperienza di Veneto Lavoro*) analizza le opportunità e le criticità legate all'impiego dei dati amministrativi per la ricerca sociale ed economica, ponendo in evidenza come tali risorse siano strettamente connesse all'evoluzione tecnologica, alle trasformazioni istituzionali e alle interpretazioni della missione da parte degli attori pubblici. In una prima parte l'autore ribadisce che i dati amministrativi rappresentano un patrimonio di informazioni a basso costo, coprente l'intera popolazione e potenzialmente conformi ai principi FAIR della European Data Strategy – per la quale i dati devono essere *Findable, Accessible, Interoperable and Reusable* – e che se adeguatamente trattati, possano integrare o sostituire le tradizionali indagini campionarie, migliorando la tempestività e la granularità delle analisi a supporto delle politiche pubbliche. Tuttavia, la loro fruibilità è condizionata dai vincoli imposti dalla normativa sulla protezione dei dati personali, ed anche da 'gelosie' proprietarie delle amministrazioni. Il documento prosegue con una panoramica storica del mercato del lavoro italiano, evidenziando come, fino agli anni Ottanta, le statistiche fossero quasi esclusivamente basate sull'indagine campionaria Istat, con limitate capacità di disaggregazione territoriale e settoriale. Parallelamente, le attività amministrative – ad esempio le Comunicazioni Obbligatorie (CO) delle imprese e le dichiarazioni dei disoccupati nei Centri per l'impiego – producevano volumi di dati che, pur essendo poco valorizzati per costi e complessità operative, avrebbero potuto fornire indicatori più tempestivi e granulari. In questo contesto Anastasia ricostruisce il percorso dell'Osservatorio di Veneto Lavoro, con la sperimentazione dell'uso dei microdati provenienti dal software Netlabor e da successive versioni telematiche delle stesse Comunicazioni Obbligatorie. I progressi tecnologici, a partire dalla digitalizzazione delle fonti cartacee degli anni Novanta fino all'introduzione delle CO telematiche nel 2008, hanno consentito la costruzione di statistiche di flusso, più adatte all'analisi longitudinale rispetto alle tradizionali misurazioni di stock, permettendo un monitoraggio mensile dei flussi occupazionali, con tempi di rilascio inferiori a un mese, migliorando così la capacità di risposta a crisi come la recessione del 2008 e la pandemia da Covid-19. Vengono inoltre esposti casi di utilizzo dei dati amministrativi per analizzare contratti a termine, misure di incentivazione all'assunzione, profili dei lavoratori immigrati e percorsi di inserimento dei giovani e dei disabili. Tuttavia, la frammentazione delle infrastrutture regionali, le divergenze di qualità e l'assenza di una governance coordinata hanno ostacolato la realizzazione di basi dati stabili a livello regionale. Anastasia evidenzia infine l'esistenza di altre barriere residue: la mancanza di standard comuni tra le amministrazioni, la difficoltà di integrazione con altri archivi (ad esempio i dati fiscali) e la riluttanza delle istituzioni a condividere informazioni per timori di privacy e di valutazione delle proprie performance operative. In conclusione, Anastasia suggerisce che, affinché i dati amministrativi diventino un vero bene pubblico per la ricerca, sia necessario un investimento continuo in infrastrutture sicure, in regole di accesso trasparenti e in pratiche di pseudonimizzazione, nonché un mandato chiaro per la produzione statistica indipendente.

Con **Buonocore, Concetti, D'Alessandro, Fidanza, Pantoni e Radini** il numero monografico approfondisce una delle caratteristiche che rende particolarmente vantaggioso l'uso dei dati amministrativi per la ricerca, vale a dire la loro interoperabilità attraverso fonti diverse. Il loro contributo (*Perché essere interoperabili. Ontologie, vocabolari controllati e schemi di dati del Catalogo nazionale dati*) analizza la necessità di rendere interoperabili i dati amministrativi al fine di trasformarli da risorsa operativa a bene pubblico capace di alimentare politiche basate su evidenze. Gli autori inquadrano la questione nel contesto della European Data Strategy, evidenziando come i principi FAIR costituiscano la base per un mercato unico dei dati anche a livello nazionale. Il testo prosegue descrivendo la frammentazione attuale dei sistemi informativi pubblici in Italia: su quasi 12.000 enti rispondenti al censimento delle istituzioni pubbliche, solo il 15,8% detiene una banca dati interoperabile, il 24,7% potrebbe renderla tale. Questa dispersione ostacola l'integrazione, la qualità delle analisi e l'adozione di tecnologie avanzate come l'Intelligenza artificiale. A fronte di tali criticità, gli autori propongono l'interoperabilità come requisito fondamentale, articolata in quattro livelli – giuridico, organizzativo, semantico, tecnico – secondo il quadro europeo. Nel livello semantico, l'articolo si concentra sugli strumenti messi a disposizione dal Catalogo nazionale dati (schema.gov.it), gestito da Istat, che

raccoglie ontologie, vocabolari controllati e schemi di dati. Queste risorse, identificate da URI (*Uniform Resource Identifier*) persistenti e pubblicate in formati aperti (OWL, RDF, JSONLD), consentono di standardizzare concetti quali persone, organizzazioni, luoghi e servizi, facilitando la condivisione e il riuso. L'iniziativa è integrata nella Piattaforma digitale nazionale dati (PDND) e nel principio *once only*, che permette ad un'amministrazione di condividere dati già in possesso con altre entità autorizzate, riducendo duplicazioni e costi. Il contributo ne evidenzia inoltre le implicazioni per la politica pubblica: dati interoperabili migliorano la tempestività e la precisione delle decisioni, favoriscono valutazioni di impatto più robuste, aumentano la trasparenza e la responsabilità amministrativa e aprono la strada a servizi digitali innovativi basati su IA. Tuttavia, gli autori avvertono che la piena valorizzazione richiede governance condivisa, risorse infrastrutturali, competenze tecniche e il rispetto rigoroso del GDPR. In conclusione, sostengono gli autori, la standardizzazione semantica e l'interoperabilità dei dati amministrativi rappresentano un fattore abilitante cruciale per trasformare il patrimonio informativo pubblico in un'infrastruttura di conoscenza capace di sostenere una Pubblica amministrazione più efficiente, trasparente e orientata al valore pubblico.

Con il contributo di **Marinuzzi e Tortorella** (*Open data e orientamento delle politiche pubbliche nella politica di coesione. Dall'accountability al policy learning*) si evidenzia la capacità dei dati amministrativi di essere, oltre che uno strumento di trasparenza informativa, anche una vera infrastruttura di orientamento delle politiche. Il caso di studio è offerto dagli open data nella politica di coesione europea. Marinuzzi e Tortorella osservano che, sebbene gli open data fossero tradizionalmente intesi come un 'addon' reputazionale, nel periodo di programmazione 2021-2027 – e ancor più in quello successivo 2028-2034 – la flessibilità, il reindirizzamento delle priorità e la rapidità di attuazione sono diventati elementi strutturali della politica di coesione. In questo contesto, la questione non è più se pubblicare i dati, ma come costruire un sistema informativo aperto capace di trasformare i dati amministrativi in capacità di definire priorità, selezionare strumenti, gestire tradeoff territoriali, correggere in corso d'opera le azioni programmatiche e favorire l'apprendimento comparativo tra programmi e territori. Il caso italiano è illustrato attraverso il passaggio dagli elenchi dei beneficiari pubblicati dalle singole Autorità di Gestione fino allo sviluppo dell'iniziativa OpenCoesione, che dal 2012 mette a disposizione dataset dettagliati su programmazione e attuazione dei fondi di coesione. L'articolo ricostruisce le difficoltà dei primi cicli (2007-2013) – distribuzione eterogenea degli elenchi dei beneficiari, formati non standardizzati, assenza di licenze di riuso e scarsa qualità dei file PDF – e mostra come il Regolamento generale sui Fondi strutturali 2014-2020 abbia introdotto l'obbligo di pubblicare le informazioni in formato aperto, con licenze di riuso e metadati più ricchi. La successiva normativa sui fondi 2021-2027 ha consolidato OpenCoesione come portale unico, garantendo l'accessibilità dei dati in formato CSV, aggiornamenti periodici e la possibilità di scaricare l'intera base dati a livello di singolo progetto. Marinuzzi e Tortorella sottolineano che la mera disponibilità di dataset non genera impatto automatico. La trasparenza funzionale, o 'targeted transparency', è necessaria per rendere le informazioni comparabili e azionabili. In questo senso, gli open data supportano tre funzioni proprie delle politiche pubbliche: l'allocazione informata delle risorse, l'adattamento in itinere delle misure e l'apprendimento comparativo. Il contributo dimostra, analizzando i Fondi FESR e FSE+ 2021-2027, come la granularità dei dati di OpenCoesione consenta di rilevare pattern di spesa, identificare colli di bottiglia, valutare la distribuzione territoriale dei beneficiari e monitorare la microprogettualità, soprattutto nel FSE+. Infine, l'articolo descrive l'interoperabilità di OpenCoesione con altre banche dati quali la Banca dati delle Amministrazioni pubbliche BDAP-MEF, ReGIS e Italia Domani, attraverso identificativi univoci comuni. Tale integrazione permette di incrociare informazioni su soggetti beneficiari, localizzazione e tipologia di finanziamento, rendendo possibile un'analisi più robusta dei flussi di risorse e delle loro conseguenze sui territori. Marinuzzi e Tortorella ne concludono che OpenCoesione è una risorsa capace di guidare le decisioni politiche, facilitare la riallocazione delle risorse e supportare un apprendimento comparato a livello europeo. Tuttavia, avvertono che il pieno utilizzo di questi dati richiede ancora il rispetto di standard di metadati, la continuità della pubblicazione e l'integrazione strutturata con i cicli decisionali, elementi imprescindibili per trasformare i dati amministrativi in bene pubblico efficace.

Il contributo di **De Paola, Paiella e Santoro** (*VisitINPS e l'uso dei dati amministrativi per la conoscenza e la valutazione delle politiche pubbliche*) si focalizza su quella che è, ad oggi, probabilmente

l'esperienza più nota di accesso strutturato agli archivi amministrativi italiani a scopo di analisi statistica, ovvero quella maturata all'interno dell'Istituto nazionale della previdenza sociale (Inps) soprattutto attraverso i programmi VisitINPS e più recentemente VisitPNRR-Lavoro. Gli archivi dell'Inps includono dati contributivi, informazioni su pensioni, prestazioni di welfare e registri aziendali, e coprono circa 25 milioni di lavoratori, 1,7 milioni di imprese e 16 milioni di pensionati. Questi dataset garantiscono copertura totale della popolazione, continuità temporale e accuratezza derivante da registrazioni legali, consentendo analisi longitudinali e studi su gruppi poco numerosi. L'accesso avviene attraverso quattro canali: osservatori statistici (dati aggregati gratuiti), open data (dataset anonimizzati scaricabili), fornitura dati su richiesta (dataset personalizzati a pagamento per soggetti non Sistan) e la piattaforma VisitINPS, che fornisce microdati ad alta granularità in ambienti protetti, previa selezione mediante bandi competitivi. Il programma VisitINPS ha finora supportato più di 260 progetti, generato 94 WorkINPS Papers e organizzato otto convegni, dimostrando un forte interesse da parte della comunità scientifica. Un'estensione più recente, VisitPNRR-Lavoro, sfrutta le stesse infrastrutture per valutare le misure del Piano nazionale di ripresa e resilienza, integrando dati Inps con quelli della Ragioneria generale dello Stato. De Paola, Paiella e Santoro individuano quattro linee di sviluppo futuro: ampliamento dell'accesso, maggiore interoperabilità tra archivi, integrazione con survey mirate e diffusione di microdati anonimizzati o sintetici. Si sottolinea la necessità di potenziare le risorse amministrative e tecniche, migliorare la documentazione dei dataset e offrire percorsi formativi per ricercatori, in particolare per giovani studiosi. L'articolo conclude che, sebbene vi siano criticità operative e di governance, l'esperienza VisitINPS rappresenta, almeno in Italia, un modello per l'accesso sicuro ai dati, capace di generare evidenze empiriche utili al policy making e di promuovere una cultura dell'evidence-based policy.

Con il contributo di **Brunetti, Filippi, Gualtieri e Ricci** (*Istruzione, salari ed eterogeneità delle imprese: evidenze da dati employer-employee linked*) il potenziale degli archivi longitudinali dell'Inps, in continuità con il contributo di De Paola *et al.* sopra richiamato, viene ulteriormente evidenziato attraverso l'integrazione con le informazioni provenienti dalle Comunicazioni Obbligatorie nazionali – che Inapp prepara a scopo di analisi statistica su mandato del Ministero del Lavoro – e con dettagli qualitativi e organizzativi provenienti da una fonte di tipo survey come la RIL (Rilevazione Imprese e Lavoratori), un'indagine campionaria sulle imprese operanti in Italia che Inapp produce a partire dal 2005, arrivata nel 2025 alla settima edizione, ribadendo con ciò come l'interoperabilità delle banche dati permetta di evitare la duplicazione di informazioni quando queste possono essere più convenientemente integrate attraverso fonti diverse. Questa integrazione permette infatti di associare a ciascun lavoratore non solo la propria storia occupazionale e retributiva, ma anche le caratteristiche di governance, di relazioni industriali, di proprietà familiare e di appartenenza a reti datoriali dell'impresa in cui lavora, informazioni tipicamente assenti nelle fonti di tipo amministrativo. Il risultato è una base di conoscenza più strutturata e potente, in grado di supportare sia la ricerca accademica sia la definizione di politiche del lavoro più mirate, evidenziando il potenziale di un approccio basato su dati integrati per comprendere la relazione tra istruzione e mercato del lavoro. Per dimostrare questo potenziale, autori e autrici si chiedono come le caratteristiche strutturali delle imprese possano modellare i rendimenti salariali dell'istruzione universitaria, concentrandosi sui lavoratori laureati. Nello specifico, l'analisi intende indagare in che misura la differenza di retribuzione associata al titolo di studio sia riconducibile alle specifiche politiche salariali delle imprese, oltre che alle caratteristiche individuali dei lavoratori. I risultati principali mostrano che le politiche aziendali influenzano in modo positivo, seppur debole, l'evoluzione dei premi salariali dei laureati, essendo questi ultimi riconducibili prevalentemente alla componente legata alle eterogeneità e abilità degli occupati. Le evidenze illustrano, poi, come la performance competitiva – misurata dal fatturato per addetto – non si trasferisce automaticamente in un incremento dei premi salariali per i laureati e quando ciò accade avviene solo parzialmente e nel medio periodo. In sintesi, le politiche retributive aziendali costituiscono un canale mediante il quale il capitale umano dei lavoratori può essere (o meno) valorizzato, ma la misura in cui questo canale contribuisce a spiegare l'evoluzione dei rendimenti salariali dell'istruzione universitaria dipende in modo rilevante dall'organizzazione dei mercati interni e dalle specifiche caratteristiche delle imprese.

Rimanendo nell'alveo degli studi sull'istruzione, **Priulla, Contini, Della Monica, Ferrara** (*La mobilità*

regionale degli studenti universitari. Opportunità e prime evidenze da una sperimentazione di linkage tra Anagrafe nazionale studenti e banca dati WHIP) mirano a valutare la fattibilità e il valore aggiunto dell'integrazione fra l'Anagrafe nazionale degli studenti (ANS) e la banca dati WHIP (Work Histories Italian Panel) al fine di ricostruire in maniera longitudinale i percorsi di studio e lavoro degli individui, con particolare attenzione alla mobilità regionale di laureati e di studenti che abbandonano gli studi. L'obiettivo è dimostrare come tale integrazione possa superare le limitazioni dei dataset di tipo survey come AlmaLaurea, consentendo una più accurata analisi delle dinamiche di selezione territoriale e delle disuguaglianze regionali. L'ANS è un archivio amministrativo gestito dal Ministero dell'Università e della ricerca che registra, per tutti gli studenti iscritti alle università italiane (laureati e non), dati anagrafici, informazioni sulla scuola superiore, cronologia annuale di immatricolazione, corsi, crediti, cambi di Ateneo e data di laurea. Le sue potenzialità risiedono nella copertura totale della popolazione studentesca e nella granularità temporale che permette di monitorare percorsi accademici complessi. L'accesso ai dati è regolato da un protocollo di ricerca con il MUR, con restrizioni di utilizzo e protezione dei dati personali. WHIP, gestita dal Ministero della Salute, è una banca dati che ricostruisce le biografie lavorative e di salute di circa il 7% dei lavoratori italiani mediante il collegamento di Inps, Inail, Ministero della Salute e Comunicazioni Obbligatorie. Offre dati longitudinali dettagliati su contratti, regioni di lavoro, settore e professione, ma è limitata ai soggetti che hanno presentato una dichiarazione ISEE. L'accesso a WHIP avviene all'interno del Programma statistico nazionale, tramite l'Università di Torino che è tra i partecipanti della relativa scheda del PSN. L'abbinamento tra le fonti in questo caso è di tipo probabilistico ed è basato su sesso, data e Comune di nascita. Le indagini sulla mobilità territoriale mostrano che i laureati, soprattutto a livello magistrale, hanno una propensione al trasferimento verso il Nord-Ovest e verso aree di lavoro più qualificate, con tassi di mobilità del 30% (magistrali) e 23% (triennali). Gli studenti che abbandonano gli studi invece rimangono prevalentemente nella macroarea dell'Ateneo, con tassi di ritorno alla zona di origine superiori (circa 4550% contro 3035% per i laureati). Priulla *et al.* sottolineano che l'integrazione ANS-WHIP consente di analizzare con maggiore precisione le traiettorie studi-lavoro, fornendo una piattaforma metodologica per studiare le disuguaglianze regionali e valutare politiche di istruzione e lavoro. Raccomandano tuttavia lo sviluppo di infrastrutture di *linkage* più stabili, in ossequio al principio di piena interoperabilità più volte sottolineato.

Un ulteriore caso di *linkage* tra dati amministrativi privi di chiavi univoche comuni è quello presentato da **Salamone** (*Integrazione dei brevetti e delle imprese italiane*). Il suo contributo intende dimostrare come l'integrazione dei dati brevettuali con le informazioni sulle imprese possa fornire una base empirica robusta per la progettazione, l'attuazione e la valutazione delle politiche nazionali di ricerca-sviluppo, trasferimento tecnologico e industriale. Le due fonti principali sono lo European Patent Office (EPO) e i registri statistici delle imprese di fonte Istat (l'Archivio statistico delle imprese attive ASIA e il relativo Registro delle unità locali). L'EPO fornisce i record delle domande di brevetto e delle concessioni dal 2015 al 2023, includendo informazioni su inventori, richiedenti, data di deposito, classificazione IPC (International Patent Classification) e citazioni. Questi dati sono completi, standardizzati e resi pubblici tramite il server di pubblicazione europeo, ma presentano limiti legati al fatto che sono osservate solo le invenzioni brevettate, escludendo innovazioni tacite o protette da segreto industriale. ASIA, ASIA-unità locali e Frame SBS contengono dati anagrafici, di localizzazione (ragione sociale, Comune, CAP, indirizzo) ed economici (dimensione, settore, produttività) delle imprese attive. Essi garantiscono una copertura pressoché totale del panorama aziendale privato non agricolo e permettono analisi longitudinali. L'accesso ad entrambe le fonti è regolato da procedure istituzionali: i dati EPO possono essere acquisiti dal European Publication Server, mentre i dataset Istat sono disponibili per la ricerca previa autorizzazione. I dati in questione, tuttavia, non contengono una chiave di identificazione univoca in comune. Utilizzando pertanto un *record linkage* basato su ragione sociale, Comune, CAP e indirizzo, l'autore costruisce un panel di 16.492 richiedenti, di cui 11.892 sono imprese, e ottiene una copertura del 95% delle domande di brevetto (32.133 record). Le imprese rappresentano il 72% degli applicant e l'89% delle domande, con una predominanza di imprese che presentano un unico brevetto (62,6%), mentre il 7,8% mantiene una propensione persistente alla brevettazione. La dimensione aziendale è una determinante chiave: le imprese con più di 250 dipendenti forniscono quasi la metà dei brevetti, mentre la quota delle micro-imprese con brevetti è in crescita dal 21% al 30% nel periodo analizzato.

Dal punto di vista settoriale, la maggior parte dei brevetti è concentrata nel manifatturiero, ma i settori emergenti – micro-nanotecnologie, semiconduttori, intelligenza artificiale e tecnologie ambientali – mostrano i più alti tassi di variazione, suggerendo aree di intervento prioritario per le politiche di transizione digitale e verde. Le imprese brevettanti hanno una produttività del lavoro media di circa 116.000 euro per dipendente, quasi tre volte quella delle imprese non brevettanti, indicando una forte correlazione tra attività inventiva codificata e performance economica.

L'articolo di **Andresciani** (*Dagli aggregati ai microdata. Spesa sanitaria privata e incentivi fiscali in Italia*) ci mostra invece quali criticità sorgono quando non è possibile accedere a un archivio di microdati di fonte amministrativa, e si deve ricorrere a una pluralità di fonti informative, ciascuna caratterizzata da specifici punti di forza e limiti. L'obiettivo dell'autrice è fornire una valutazione critica delle principali sfide legate alla misurazione della spesa sanitaria privata, della copertura assicurativa volontaria e delle detrazioni fiscali in Italia, evidenziando come la frammentazione delle informazioni, l'eterogeneità tra le fonti e, in alcuni casi, l'elevato livello di aggregazione dei dati possano ostacolare la capacità di monitorare l'equità e l'efficacia delle politiche di finanziamento sanitario. Le banche dati esaminate comprendono i conti nazionali dei consumi classificati secondo lo scopo (c.d., COICOP - Classification of Individual CONsumption by Purpose), il Sistema di Conti sanitari, l'indagine sui consumi delle famiglie dell'Istat e i dati amministrativi del Sistema Tessera sanitaria (STS). I conti nazionali e il Sistema dei Conti sanitari forniscono stime aggregate della spesa privata, ma differiscono per schema contabile, perimetro di riferimento e inclusione di ticket o agevolazioni fiscali, rendendo complessi i confronti tra fonti. L'indagine Istat sui consumi delle famiglie offre dati a livello nazionale e regionale, ma è soggetta a variabilità campionaria e possibile sottodichiarazione. Il STS, gestito dal Ministero dell'Economia e delle finanze, registra le transazioni sanitarie a livello individuale, ma le pubblicazioni attuali forniscono solo informazioni aggregate, senza distinguere tra spesa 'out of pocket' (OOP), rimborsi assicurativi o ticket, e limitando l'analisi distributiva. In modo non sorprendente, i risultati principali mostrano quindi che (i) le stime della spesa sanitaria privata variano sensibilmente a seconda della fonte; (ii) la mancanza di microdati impedisce di distinguere chi sostiene la spesa, in che misura e con quali coperture assicurative, ostacolando analisi di equità territoriale e socioeconomica; (iii) le detrazioni fiscali sono riportate in forma aggregata, senza consentire di identificare se i benefici ricadono sui ticket del SSN o su spese interamente private, compromettendo la valutazione della loro progressività. Andresciani evidenzia che, se il STS fosse reso accessibile a livello micro per la ricerca con adeguate tutele di privacy, sarebbe possibile scomporre la spesa privata in pagamenti OOP, rimborsi assicurativi e ticket, e valutare la distribuzione degli incentivi fiscali per gruppi di reddito e regioni. Il rafforzamento della governance dei dati sulla spesa sanitaria e l'accesso a set più granulari appaiono pertanto prerequisiti imprescindibili per monitorare efficacemente il ruolo del finanziamento privato nei sistemi sanitari ibridi, per verificare l'effettivo 'effetto decongestione' da parte del sistema sanitario privato e per garantire che le politiche fiscali non generino effetti regressivi.

Gli ultimi due contributi di questo numero speciale di Sinapsi esemplificano infine il potenziale delle fonti amministrative anche per la cosiddetta analisi ex ante. L'obiettivo dell'articolo di **Bosia, Carlone, Floredan e Prever** (*La profilazione delle persone disponibili al lavoro in Piemonte. Il ruolo dei dati amministrativi e dei dati territoriali locali*) è valutare se l'integrazione di dati amministrativi individuali con informazioni territoriali dei Sistemi locali del Lavoro (SLL) possa migliorare la capacità predittiva dei modelli di profilazione utilizzati nei servizi per l'impiego, con particolare attenzione al caso piemontese. I dati amministrativi impiegati derivano dalla Scheda anagrafico-professionale (SAP), un fascicolo digitale che aggrega informazioni anagrafiche, formative, occupazionali e di partecipazione alle politiche attive, alimentato automaticamente dalle Comunicazioni Obbligatorie e dai Centri per l'impiego. La SAP consente di osservare l'intera popolazione di soggetti che sottoscrivono la Dichiarazione di Immediata Disponibilità (DID), garantendo copertura completa, granularità elevata e aggiornamento continuo. Per arricchire l'analisi, autori e autrici attingono al Sistema informativo dei SLL, prodotto da Istat, che combina registri delle imprese (ASIA), del lavoro (con il Registro base degli Individui, RBI) e dell'anagrafe per costruire indicatori territoriali. Queste variabili forniscono una descrizione dettagliata della struttura produttiva e demografica dei territori locali. I risultati principali mostrano che il modello replicato a livello regionale riproduce molte delle regolarità individuate dal modello nazionale, ma con

differenze marcate: l'esperienza lavorativa recente ha un impatto preventivo sulla non occupazione più forte in Piemonte, mentre l'età e il genere risultano meno determinanti. L'inclusione di variabili a livello di SLL riduce l'errore di previsione, evidenziando che un contesto con elevata quota di addetti stranieri aumenta il rischio di non reinserimento, mentre una più alta presenza di lavoratori indipendenti è associata a un minor rischio, a patto che l'autoimpiego sia strutturale. I risultati suggeriscono inoltre che l'impiego di algoritmi più flessibili può individuare più efficacemente i soggetti a più alto rischio di esclusione prolungata dal mercato del lavoro.

Arriviamo così all'ultimo contributo incluso in questo numero speciale, quello di **Bazzoli, Marzadro, Tassi e Fiorio** (*TREMODO-Imprese: un modello di microsimulazione*), il cui obiettivo è quello di presentare TREMOD-Imprese, un modello di microsimulazione fiscale che consente di valutare ex ante le politiche di fiscalità d'impresa a livello della Provincia autonoma di Trento. TREMOD-Imprese si basa su un set di dati amministrativi costituiti dalle dichiarazioni IRES, IRAP e IVA dell'anno fiscale 2019, integrati con l'Archivio Istat delle imprese attive (ASIA). I dati fiscali forniscono informazioni dettagliate su utili, valore della produzione, volume d'affari, aliquote, detrazioni e crediti d'imposta per ciascuna impresa, mentre l'anagrafica consente di ricavare dimensione (classi di addetti), settore, sede legale e altri caratteri strutturali. L'integrazione avviene mediante il codice fiscale dell'impresa, garantendo una corrispondenza univoca tra le tre dichiarazioni e l'archivio ASIA. L'accesso a tali banche dati è reso possibile dall'accordo attuativo Tremod tra l'Istituto di Statistica della Provincia di Trento (ISPAT) e la Fondazione Bruno Kessler, rientrante nel Protocollo d'intesa per l'attuazione di ricerche statistiche, e che assicura la sicurezza delle informazioni e il rispetto della riservatezza. Le simulazioni esemplificative dimostrano che la rimozione delle detrazioni IRES comporterebbe un incremento di 1,6 milioni di euro (circa < 1% del gettito IRES), mentre l'eliminazione delle aliquote ridotte porterebbe a un aumento di 5,2 milioni di euro ($\approx$ 2%). Per l'IRAP, l'aumento dell'aliquota standard al livello nazionale (da 2,68% a 3,9%) produrrebbe un incremento di circa 33 milioni di euro ($\approx$ 33% del gettito IRAP), mentre l'eliminazione delle deduzioni regionali genererebbe solo un aumento marginale di 1,1 milioni di euro. Le simulazioni sono statiche, non prevedono reazioni comportamentali delle imprese e sono limitate dalla mancanza di dati su crediti d'imposta specifici e su alcune variabili necessarie per scenari più complessi. Autori e autrici concludono che l'integrazione di dati fiscali amministrativi offre un valido strumento per la valutazione ex ante di politiche fiscali anche a livello subnazionale, ma evidenziano la necessità di ampliare la copertura temporale e di arricchire il set di variabili per consentire simulazioni più sofisticate e attendibili, per esempio per includere reazioni di tipo comportamentale.

3. Questioni aperte

I contributi raccolti in questo numero di *Sinappsi* mostrano con chiarezza il potenziale dei dati amministrativi per la produzione di evidenza empirica e per il miglioramento delle politiche pubbliche, ma mettono anche in evidenza alcuni ostacoli che in Italia ne limitano ancora la piena valorizzazione. Le questioni aperte non riguardano soltanto la disponibilità materiale dei dati: investono le condizioni giuridiche e organizzative dell'accesso, la capacità di integrare fonti differenti, la qualità e la documentazione degli archivi, nonché le infrastrutture necessarie a renderli effettivamente utilizzabili.

1. **Governance, coordinamento e regole comuni.** Autori e autrici evidenziano innanzitutto la frammentazione dei sistemi informativi e l'assenza, in alcuni ambiti, di una governance coordinata. L'esperienza dei dati sul mercato del lavoro descritta da Anastasia, per esempio, mostra divergenze nella qualità delle infrastrutture regionali, mancanza di standard comuni e difficoltà di costruire basi dati stabili. Analogamente, la condivisione delle informazioni può essere ostacolata dalla riluttanza delle istituzioni a rendere disponibili i propri archivi. Secondo gli autori e le autrici dei contributi pubblicati in questo fascicolo, il pieno utilizzo dei dati amministrativi richiede quindi regole di accesso trasparenti e un coordinamento capace di favorire la condivisione.
2. **Tutela dei dati personali e accesso alla ricerca.** Un secondo ostacolo riguarda la necessità di conciliare l'utilizzo dei dati per finalità statistiche e di ricerca con la protezione delle informazioni personali. Il quadro normativo descritto da di Somma consente il riutilizzo dei dati, ma lo sottopone a condizioni di liceità, garanzie e limiti, mentre l'accesso ai microdati richiede procedure e misure di protezione specifiche, quali la pseudonimizzazione, anonimizzazione, ambienti protetti e monitoraggio delle

pratiche di accesso. La questione aperta consiste dunque nel rendere accessibili i dati alla ricerca, nel quadro della disciplina sulla protezione dei dati personali, in modo più sistematico e trasparente di quanto accada oggi. Tutto ciò è possibile, come dimostrato dai numerosi contributi pervenuti per questo numero speciale e da altre esperienze italiane e in altri Paesi dell'Unione europea soggetti ai medesimi vincoli istituzionali vigenti in Italia (Berton e Paruolo 2026).

3. **Interoperabilità, linkage e standardizzazione.** La frammentazione delle banche dati rende centrale la capacità di collegare informazioni provenienti da fonti diverse. I contributi dedicati all'interoperabilità e ai record *linkage* mostrano però che questo processo non è automatico: le fonti possono non condividere chiavi identificative univoche (Priulla *et al.*; Salamone) e possono differire per struttura, definizioni e caratteristiche informative. Buonocore *et al.* sottolineano quindi l'importanza della standardizzazione semantica, dei vocabolari controllati e degli schemi di dati, oltre che di infrastrutture di *linkage* più stabili.
4. **Qualità, documentazione e granularità dei dati.** Un ulteriore problema riguarda la possibilità di disporre di informazioni sufficientemente dettagliate e documentate, vale a dire di microdati al livello della singola unità statistica. Il contributo di Andresciani sulla spesa sanitaria privata evidenzia in modo particolare i limiti derivanti dall'accesso ai soli dati aggregati: l'assenza di microdati impedisce di distinguere chi sostiene la spesa, in quale misura e con quali forme di copertura, ostacolando le analisi distributive. Altri contributi richiamano invece la necessità di migliorare la documentazione dei dataset, la qualità dei dati e la continuità delle informazioni pubblicate. La disponibilità del dato non è quindi sufficiente se la sua granularità, qualità o documentazione non consentono un uso analitico adeguato.
5. **Infrastrutture, risorse e competenze.** Il pieno sviluppo dell'analisi statistica richiede inoltre investimenti continuativi in capitale fisico (infrastrutture sicure) e umano (risorse tecniche e organizzative da parte delle istituzioni proprietarie dei dati ma anche percorsi formativi per i ricercatori). Anche l'esperienza maggiormente strutturata all'interno del panorama italiano, ovvero quella di VisitINPS (De Paola *et al.*), indica come linee di sviluppo l'ampliamento dell'accesso, una maggiore interoperabilità, la diffusione di microdati anonimizzati o sintetici e il rafforzamento della documentazione. La mera esistenza degli archivi amministrativi non è pertanto sufficiente; occorre la capacità di predisporre le condizioni tecniche e professionali necessarie per trasformarli in una risorsa effettivamente utilizzabile dalla ricerca.
6. **Continuità e integrazione dei dati nei processi di policy.** Infine, autori e autrici evidenziano che la disponibilità dei dati deve essere accompagnata da continuità nella pubblicazione e da un collegamento strutturato con i cicli decisionali. Il contributo di Bosia *et al.* mostra come l'integrazione di dati amministrativi con dati territoriali possa migliorare gli attuali processi decisionali che guidano l'erogazione delle politiche. L'esperienza di OpenCoesione (Marinuzzi e Tortorella), dall'altra parte, mostra che la semplice pubblicazione di dataset non produce automaticamente impatto: i dati devono essere comparabili, azionabili e disponibili con aggiornamenti periodici. A sua volta, ciò dipende da un mix di obblighi e incentivi all'utilizzo dei dati amministrativi a scopo di monitoraggio e valutazione che chiami in causa tutte le parti coinvolte: responsabili delle politiche, detentori dei dati grezzi, utilizzatori dei dati a scopo statistico.

Chiudiamo l'introduzione a questa prima parte di interventi dedicati al tema dell'utilizzo e dell'accesso alle fonti amministrative a scopo di ricerca in Italia chiedendoci perché, a fronte dei vantaggi che tali fonti offrono e che abbiamo descritto più in alto, l'Italia si trovi tutto sommato ancora a metà del guado, con tante esperienze di utilizzo interessanti dal punto di vista analitico, una tradizione piuttosto consolidata, ma un sistema nel suo complesso ancora troppo frammentato e poco strutturato per poterne sfruttare appieno tutte le potenzialità. Non riteniamo che queste ragioni siano da ricercare nella normativa a tutela della riservatezza dei dati personali e sensibili. Alcune esperienze in Italia ed altre all'interno dell'Unione europea dimostrano che, benché difficile, la costruzione di un sistema di utilizzo delle fonti amministrative a scopo di ricerca che coniughi rispetto delle norme e flessibilità nelle forme di accesso è tutt'altro che impossibile. Non riteniamo neanche che questa carenza sia da ricondurre a quelle che Berton e Paruolo (2026) etichettano come "false credenze" circa l'uso dei dati

amministrativi a scopo di ricerca. In primo luogo, circa il costo legato alla preparazione a scopo di ricerca di tali fonti: se da un lato è certamente vero che, a differenza dei dati che originano da survey, la fonte amministrativa richiede un investimento iniziale di pulizia, documentazione e preparazione dei dati grezzi, dall'altro lato tale costo è in larga misura fisso e non va sostenuto nuovamente ad ogni nuova release dei dati, a differenza di ciò che accade con le survey la cui struttura dei costi è dominata dalla componente variabile. In secondo luogo, circa i ritardi nella disponibilità del dato trattabile a scopo di analisi, non compatibili con i vincoli dell'azione politica, riteniamo invece che sia vero l'esatto contrario: il flusso di generazione del dato amministrativo grezzo è continuo per costruzione e la trasformazione in informazioni utilizzabili a scopo di analisi richiede, una volta che tali procedure siano state validate, poche settimane al massimo; viceversa i tempi di validazione ed esecuzione dei questionari sono almeno di alcuni mesi e seguono generalmente un calendario predeterminato difficilmente adattabile alle esigenze di conoscenza che emergono dalla politica.

Quelli che nel dibattito corrente vengono facilmente indicati come i 'soliti sospetti' – le norme sulla privacy, i costi, vincoli tecnici o di tempo – possono quindi spiegare solo in parte gli attuali ritardi che osserviamo in Italia. Ragioni più profonde si possono trovare estendendo il concetto di "gelosia del dato" introdotto da Anastasia, per fare riferimento a una più generale mancanza di volontà e di un adeguato sistema di incentivi che spinga le autorità preposte a investire nella direzione di un sistema coordinato e strutturato di accesso alle fonti amministrative a scopo di ricerca, con regole chiare e uniformi e sistemi di accesso che presentano costi sostenibili da parte degli utilizzatori del dato.

Bibliografia

- Berton F., Paruolo P. (a cura di) (2026), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer (forthcoming)
- Bloem M., Gorissen I., van den Boom R., van Hees V. (2026), CBS microdata: a user perspective, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 21 (forthcoming)
- Commissione europea (2026), *Review of the 2021 European Pillar of Social Rights Action Plan*, Bruxelles, SWD (2026) 650 final
- Consiglio dell'Unione europea (2024), *Draft Council Conclusions on the role of labour market, skills and social policies for resilient economies*, Bruxelles, 11066/24
- Contini B., Trivellato U. (a cura di) (2005), *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*, Bologna, il Mulino
- Crato N., Paruolo P. (a cura di) (2019), *Data-driven policy impact evaluation. How to access microdata is transforming policy design*, Berlino, Springer
- Fabre B., Gadouche K. (2026), Access to French administrative microdata, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 11 (forthcoming)
- Leombruni R., Ardito C. (2026), The long and winding road. The access to administrative data for research in Italy, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 16 (forthcoming)
- Lima F., Martins J.P., Neugebauer K., Pimentel J.F., Rodrigues S., Saraiva P., Sousa R., Zilhão M.J. (2026), Administrative data in Portugal, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 23 (forthcoming)

Using administrative data in Italy to inform policymaking

Fabio Berton

European Commission, Joint Research Centre

Roberto Leombruni

Università di Torino

1. Why administrative data?

Issue 2/2026 of the journal *Sinappsi* is the first of two special issues devoted to the state of the art in the use of administrative databases for research purposes in Italy. The aim is to highlight and illustrate the features that make administrative data particularly well suited to statistical analysis and public policy evaluation, as evidenced not only by the growing demand from the research community for access to these databases, but also by the increasing policy attention they are receiving at both Italian and European level.

By administrative data, we mean “*data collected by public entities for their service activities*” (Crato and Paruolo 2019, v); more precisely, we refer to the microdata collected by such entities in carrying out their functions, which therefore contain information at the level of the individual, household, firm, etc., rather than aggregated information. The ‘*long and winding*’ history of their use for scientific research in Italy (Leombruni and Ardito 2026) began in the mid-1980s, when Bruno Contini, now Professor Emeritus of Econometrics at the University of Turin, and Riccardo Revelli – who passed far too soon – came up with the idea of transforming the social security contribution records held by the National Social Security Institute (INPS, Italy) into individual employment histories, what we now commonly refer to as LEED (Linked Employer-Employee Data). Since then, this history has evolved considerably, both in breadth – because the fields of study covered by administrative information now extend well beyond the material contained in INPS archives – and in depth, with new forms of data access seeking, with varying degrees of success, to reconcile researchers’ flexibility of use with the protection of data confidentiality. This special issue of *Sinappsi* aims to provide as comprehensive an account as possible of the current state of the art in Italy regarding the use of administrative data sources for statistical analysis.

At least two factors explain the centrality of this issue in the current debate. On the one hand, several characteristics of administrative microdata make them particularly well suited to statistical analysis: (i) because they are collected by public administrations in the course of their activities, administrative data often cover the entire target population rather than a sample; this makes it possible to analyse the distributional effects of the policies under consideration more precisely, and therefore to understand not only whether a policy works, but also for whom it works. (ii) Unlike sample surveys, administrative data do not depend on individuals’ ability to recall past events or on their willingness to report potentially sensitive information. Basic information is collected in the course of administrative activities, thereby reducing some of the typical problems associated with measurement error, including unintentional error. (iii) Their use is also advantageous from an economic perspective: the costs of data collection have already been incurred for the original administrative purpose. (iv) In addition, the presence of unique identifiers makes it possible to link different records and, where possible, to integrate them with survey data, substantially expanding the amount of information available for analysis. (v) Finally, continuous data collection makes these data available virtually in real time and makes it possible to follow individuals, households or firms over very long periods.

On the other hand – and at least partly as a consequence of what has just been described – there is growing demand, from the research community and in some cases from public administrations themselves, for access to integrated administrative data sources in order to improve the quality and

quantity of research. In their contribution to this special issue, De Paola *et al.* report that the VisitINPS programme alone, operational since 2016, has so far supported more than 260 research projects and the publication of around 100 articles in its dedicated series. Looking beyond Italy, Lima *et al.* (2026) report a similar number of projects carried out through the Microdata Research Laboratory of the Banco de Portugal (BPLIM), involving around 350 researchers. Since 2012, the *Centre d'accès sécurisé aux données* (CASD) of the French National Institute of Statistics has hosted around 800 scientific projects involving more than 1,600 active researchers from around the world. Likewise, as of 2024, the 7,500 databases available through the Centraal Bureau voor de Statistiek (CBS) in the Netherlands had attracted the interest of around 1,900 researchers (Bloem *et al.* 2026).

The combined outcome of these developments is growing attention to administrative data not only from the research community and from the institutions and agencies that hold the underlying data, but also from policymakers: for example, in June 2024, the Council of the European Union called on the European Commission to “enhance access to administrative data for policy evaluation and impact assessment purposes” (Council of the European Union 2024, point 13). In July 2026, the European Commission itself acknowledged that “Administrative data held by public authorities are a unique source of evidence for policy making and monitoring” and that “The use of administrative data can be more efficient and less costly than existing alternatives” (European Commission 2026). We were therefore not surprised, although we were equally pleased, that the call for papers for this special issue of Sinappsi on the use of and access to administrative data in Italy received such a strong response that publication of the accepted papers had to be split across two consecutive issues. In short, paraphrasing a volume edited by Bruno Contini himself together with Ugo Trivellato (2005), ‘And yet it moves’.

The content of the special issue

The issue opens with a legal framework for the topic. **di Somma**’s contribution (*Administrative data: GDPR rules for statistical and research use*) analyses the Italian and European regulatory framework governing the processing of administrative data for statistical and research purposes, with particular reference to the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679). The author starts from the observation that the information assets held by public administrations have become strategic for evidence-based policymaking, while a large proportion of administrative databases contain personal information, in some cases sensitive data. In this context, legislation does not prevent their reuse, but strictly regulates the conditions under which such processing is lawful, the safeguards to be adopted and the limits on use. The paper reconstructs the origins of statistical processing within the National Statistical System (Sistan) and the regulations that established the National Statistical Programme and required the interconnection of public information systems. It outlines the provisions guaranteeing statistical confidentiality and prohibiting administrative repercussions, thereby preventing data collected for statistical purposes from being used for other purposes. The Privacy Code is complemented by rules of professional conduct governing statistical processing and by arrangements for external researchers’ access to data, including requirements for the anonymisation or pseudonymisation of datasets. The framework was subsequently supplemented by amendments aligning national legislation with the GDPR, introducing specific provisions for statistical activities. The GDPR, in fact, provides specific derogations from the principles of purpose limitation and storage limitation, as well as exemptions from the prohibition on processing special categories of personal data, where such processing is carried out for statistical purposes and appropriate safeguards are in place, including pseudonymisation. In conclusion, the author argues that, alongside a robust regulatory framework, effective use of administrative data requires continuous investment in technical measures (pseudonymisation, anonymisation and privacy-enhancing technologies) and in monitoring access practices. Only a careful balance between the protection of fundamental rights and research needs, di Somma argues, can ensure that these data continue to constitute a public good capable of improving evidence-based policymaking.

Once the legal framework has been established, the issue turns to the contribution of one of the historically most important voices in the use of administrative data sources for statistical purposes, Bruno **Anastasia**. His contribution (*Notes on the uses and challenges of administrative data: the*

experience of Veneto Lavoro) analyses the opportunities and critical issues associated with the use of administrative data for social and economic research, highlighting how such resources are closely linked to technological developments, institutional change and the ways in which public actors interpret their mission. In the first part, the author reiterates that administrative data constitute a low-cost information asset covering the entire population and potentially consistent with the FAIR principles of the European Data Strategy – according to which data should be Findable, Accessible, Interoperable and Reusable – and that, if properly processed, they can complement or replace traditional sample surveys, improving the timeliness and granularity of analyses supporting public policy. Their usability, however, is constrained by personal data protection legislation and also by administrations' proprietary "data ownership" concerns. The paper then provides a historical overview of the Italian labour market, showing that until the 1980s statistics were based almost exclusively on the Istat sample survey, with limited capacity for territorial and sectoral disaggregation. At the same time, administrative activities – for example, firms' Mandatory Communications (*Comunicazioni Obbligatorie*, CO) and declarations submitted by unemployed people at public employment services – generated large volumes of data which, although underused because of their costs and operational complexity, could have provided more timely and granular indicators. Against this background, Anastasia reconstructs the experience of the Veneto Lavoro Observatory, including experimentation with microdata from the Netlabor software and subsequent electronic versions of the same Mandatory Communications. Technological advances, from the digitisation of paper-based sources in the 1990s to the introduction of electronic CO in 2008, made it possible to construct flow statistics, better suited to longitudinal analysis than traditional stock measures, enabling monthly monitoring of employment flows, with release times of less than one month and thus improving the capacity to respond to crises such as the 2008 recession and the COVID-19 pandemic. The paper also presents cases in which administrative data have been used to analyse fixed-term contracts, hiring incentives, the profiles of immigrant workers, and the labour market integration pathways of young people and people with disabilities. However, fragmentation of regional infrastructures, differences in quality and the absence of coordinated governance have hindered the creation of stable regional databases. Anastasia also highlights other remaining barriers: the lack of common standards across administrations, difficulties in linking data with other archives (for example, tax data), and institutions' reluctance to share information because of concerns about privacy and the assessment of their operational performance. In conclusion, Anastasia suggests that, for administrative data to become a genuine public good for research, continuous investment is needed in secure infrastructures, transparent access rules and pseudonymisation practices, together with a clear mandate for independent statistical production.

With the contribution by **Buonocore, Concetti, D'Alessandro, Fidanza, Pantoni and Radini**, the special issue explores one of the features that makes the use of administrative data particularly advantageous for research, namely their interoperability across different sources. Their contribution (*Why be interoperable? Ontologies, controlled vocabularies and data schemas of the National Data Catalogue*) analyses the need to make administrative data interoperable in order to transform them from an operational resource into a public good capable of supporting evidence-based policymaking. The authors place the issue in the context of the European Data Strategy, highlighting how the FAIR principles provide the basis for a data single market also at national level. The paper goes on to describe the current fragmentation of public information systems in Italy: among almost 12,000 entities responding to the census of public institutions, only 15.8% hold an interoperable database, while 24.7% could make theirs interoperable. This fragmentation hampers integration, the quality of analysis and the adoption of advanced technologies such as artificial intelligence. In response to these challenges, the authors propose interoperability as a fundamental requirement, articulated at four levels – legal, organisational, semantic and technical – in line with the European framework. At the semantic level, the article focuses on the tools made available through the National Data Catalogue (schema.gov.it), managed by Istat, which brings together ontologies, controlled vocabularies and data schemas. These resources, identified through persistent URIs (Uniform Resource Identifiers) and published in open formats (OWL, RDF, JSON-LD), make it possible to standardise concepts such as people, organisations, places and services, facilitating data sharing and reuse. The initiative is integrated into the National

Digital Data Platform (PDND) and the once-only principle, which allows an administration to share data already in its possession with other authorised entities, reducing duplication and costs. The contribution also highlights the implications for public policy: interoperable data improve the timeliness and accuracy of decisions, support more robust impact assessments, increase administrative transparency and accountability, and pave the way for innovative AI-based digital services. However, the authors warn that fully realising this potential requires shared governance, infrastructure resources, technical expertise and strict compliance with the GDPR. In conclusion, they argue that semantic standardisation and interoperability of administrative data are crucial enabling factors for transforming public-sector information assets into a knowledge infrastructure capable of supporting a more efficient, transparent public administration oriented towards public value.

The contribution by **Marinuzzi** and **Tortorella** (*Open data and the orientation of public policies in cohesion policy. From accountability to policy learning*) highlights the capacity of administrative data to serve not only as a tool for information transparency but also as a genuine infrastructure for steering policy. The case study concerns open data in European cohesion policy. Marinuzzi and Tortorella observe that, although open data were traditionally viewed as a reputational 'add-on', during the 2021-2027 programming period – and even more so in the subsequent 2028-2034 period – flexibility, the redirection of priorities and speed of implementation have become structural elements of cohesion policy. In this context, the issue is no longer whether to publish data, but how to build an open information system capable of turning administrative data into the capacity to set priorities, select instruments, manage territorial trade-offs, adjust programme actions during implementation and foster comparative learning across programmes and territories. The Italian case is illustrated through the transition from beneficiary lists published by individual Managing Authorities to the development of the OpenCoesione initiative, which since 2012 has made detailed datasets on the programming and implementation of cohesion funds available. The article reconstructs the difficulties of the early programming cycles (2007-2013) – heterogeneous distribution of beneficiary lists, non-standardised formats, absence of reuse licences and poor-quality PDF files – and shows how the General Regulation on the Structural Funds for 2014-2020 introduced an obligation to publish information in open formats, with reuse licences and richer metadata. Subsequent legislation for the 2021-2027 funds consolidated OpenCoesione as a single portal, ensuring access to data in CSV format, periodic updates and the possibility of downloading the entire database at project level. Marinuzzi and Tortorella stress that the mere availability of datasets does not automatically generate impact. Functional, or "targeted", transparency is needed to make information comparable and actionable. In this sense, open data support three functions intrinsic to public policy: informed resource allocation, in-course adaptation of measures and comparative learning. Analysing the 2021-2027 ERDF and ESF+ funds, the contribution shows how the granularity of OpenCoesione data makes it possible to identify spending patterns, bottlenecks, the territorial distribution of beneficiaries and micro-project activity, especially under the ESF+. Finally, the article describes the interoperability of OpenCoesione with other databases, such as the Public Administrations Database (BDAP-MEF), ReGiS and Italia Domani, through common unique identifiers. This integration makes it possible to cross-reference information on beneficiaries, location and type of funding, enabling more robust analysis of resource flows and their consequences across territories. Marinuzzi and Tortorella conclude that OpenCoesione is a resource capable of informing policy decisions, facilitating the reallocation of resources and supporting comparative learning at European level. However, they warn that fully exploiting these data still requires compliance with metadata standards, continuity of publication and structured integration with decision-making cycles, all of which are essential for turning administrative data into an effective public good.

The contribution by **De Paola**, **Paiella** and **Santoro** (*VisitINPS and the use of administrative data for knowledge and public policy evaluation*) focuses on what is, to date, probably the best-known experience of structured access to Italian administrative archives for statistical analysis, namely the experience developed within the National Social Security Institute (INPS), especially through the VisitINPS programme and, more recently, VisitPNRR-Lavoro. INPS archives include contribution records, information on pensions, welfare benefits and business registers, and cover around 25 million workers, 1.7 million firms and 16 million pensioners. These datasets provide full population coverage, temporal

continuity and accuracy derived from legally mandated records, enabling longitudinal analyses and studies of small population groups. Access is provided through four channels: statistical observatories (free aggregated data), open data (downloadable anonymised datasets), data provision upon request (customised datasets supplied for a fee to non-Sistan entities) and the VisitINPS platform, which provides highly granular microdata in secure environments following selection through competitive calls. The VisitINPS programme has so far supported more than 260 projects, generated 94 WorkINPS Papers and organised eight conferences, demonstrating strong interest from the research community. A more recent extension, VisitPNRR-Lavoro, uses the same infrastructure to evaluate measures under the National Recovery and Resilience Plan, integrating INPS data with data from the General Accounting Department. De Paola, Paiella and Santoro identify four areas for future development: broader access, greater interoperability across archives, integration with targeted surveys, and wider availability of anonymised or synthetic microdata. They stress the need to strengthen administrative and technical resources, improve dataset documentation and provide training pathways for researchers, particularly early-career scholars. The article concludes that, although operational and governance challenges remain, the VisitINPS experience represents, at least in Italy, a model for secure data access, capable of generating empirical evidence useful for policymaking and promoting a culture of evidence-based policy.

The contribution by **Brunetti, Filippi, Gualtieri and Ricci** (*Education, wages and firm heterogeneity: evidence from linked employer-employee data*) further highlights the potential of INPS longitudinal archives, building on the contribution by De Paola *et al.* mentioned above, through integration with information from the national Mandatory Communications – which INAPP prepares for statistical analysis under a mandate from the Ministry of Labour – and with qualitative and organisational information from a survey source such as RIL (Survey on Enterprises and Workers), a sample survey of firms operating in Italy that INAPP has conducted since 2005 and which reached its seventh edition in 2025. This once again shows how interoperability across databases can avoid duplication of information when the same information can be more efficiently integrated through different sources. This integration makes it possible to associate each worker not only with their employment and earnings history, but also with characteristics relating to the governance, industrial relations, family ownership and membership of employers' networks of the firm in which they work, information typically absent from administrative sources. The result is a more structured and powerful knowledge base, capable of supporting both academic research and the design of more targeted labour policies, highlighting the potential of an integrated-data approach to understanding the relationship between education and the labour market. To demonstrate this potential, the authors ask how firms' structural characteristics may shape the wage returns to higher education, focusing on university graduates. Specifically, the analysis examines the extent to which the earnings differential associated with educational attainment can be attributed to firms' specific wage policies, in addition to workers' individual characteristics. The main results show that firm policies have a positive, albeit weak, effect on the evolution of graduates' wage premiums, which are predominantly attributable to heterogeneity and skill differences among workers. The evidence also shows that competitive performance – measured by turnover per employee – does not automatically translate into higher wage premiums for graduates and, when it does, the effect emerges only partially and over the medium term. In short, firms' pay policies constitute one channel through which workers' human capital may or may not be rewarded, but the extent to which this channel helps explain the evolution of returns to higher education depends substantially on the organisation of internal labour markets and on firms' specific characteristics.

Staying within the field of education research, **Priulla, Contini, Della Monica and Ferrara** (*Regional mobility of university students. Opportunities and preliminary evidence from a linkage procedure between the National student registry and the work histories Italian Panel*) assess the feasibility and added value of integrating the National Student Registry (ANS) with the WHIP database (Work Histories Italian Panel) in order to reconstruct individuals' study and work trajectories longitudinally, with particular attention to the regional mobility of graduates and students who drop out. The aim is to show how such integration can overcome the limitations of survey datasets such as AlmaLaurea, allowing more accurate analysis of territorial selection dynamics and regional inequalities. ANS is

an administrative archive managed by the Ministry of University and Research that records, for all students enrolled in Italian universities (graduates and non-graduates), personal data, information on upper-secondary education, annual histories of enrolment, degree programmes, credits, university transfers and graduation dates. Its potential lies in full coverage of the student population and in its temporal granularity, which makes it possible to monitor complex academic trajectories. Access to the data is governed by a research protocol with the Ministry of University and Research, with restrictions on use and protection of personal data. WHIP, managed by the Ministry of Health, is a database that reconstructs the employment and health histories of around 7% of Italian workers by linking INPS, INAIL, Ministry of Health, and Mandatory Communications data. It provides detailed longitudinal information on contracts, regions of employment, sector and occupation, but is limited to individuals who have submitted an ISEE declaration. Access to WHIP takes place within the National Statistical Programme, through the University of Turin, which is one of the participants in the relevant NSI programme entry. The linkage between the sources is probabilistic and is based on sex, date and municipality of birth. The findings on geographical mobility show that graduates, especially those holding a master's degree, tend to move towards north-western Italy and towards areas with more highly qualified employment, with mobility rates of 30% for master's graduates and 23% for bachelor's graduates. Students who drop out, by contrast, tend to remain in the macro-area of their university, with higher return-to-origin rates (around 45-50% compared with 30-35% for graduates). Priulla *et al.* stress that ANS-WHIP integration makes it possible to analyse study-to-work trajectories more accurately, providing a methodological platform for studying regional inequalities and evaluating education and labour policies. They nevertheless recommend developing more stable linkage infrastructures, in line with the principle of full interoperability repeatedly emphasised throughout the issue.

A further case of linkage between administrative data lacking common unique identifiers is presented by **Salamone** (*Linking patents and Italian firms*). His contribution aims to show how integrating patent data with firm information can provide a robust empirical basis for the design, implementation and evaluation of national policies on research and development, technology transfer and industrial policy. The two main sources are the European Patent Office (EPO) and Istat statistical business registers (the Statistical Register of Active Enterprises, ASIA, and the related local-unit register). The EPO provides records of patent applications and grants from 2015 to 2023, including information on inventors, applicants, filing date, IPC (International Patent Classification) and citations. These data are complete, standardised and made publicly available through the European publication server, but they have limitations because they cover only patented inventions, excluding tacit innovations or innovations protected by trade secrecy. ASIA, ASIA-local units and Frame SBS contain demographic, location (company name, municipality, postcode, address) and economic data (size, sector, productivity) on active firms. They provide almost complete coverage of the non-agricultural private business population and allow longitudinal analyses. Access to both sources is governed by institutional procedures: EPO data can be obtained from the European Publication Server, while Istat datasets are available for research subject to authorisation. The sources, however, do not contain a common unique identifier. The author therefore uses record linkage based on company name, municipality, postcode and address to construct a panel of 16,492 applicants, of whom 11,892 are firms, achieving coverage of 95% of patent applications (32,133 records). Firms account for 72% of applicants and 89% of applications, with a predominance of firms filing a single patent (62.6%), while 7.8% show a persistent propensity to patent. Firm size is a key determinant: firms with more than 250 employees account for almost half of all patents, while the share of micro-firms with patents increased from 21% to 30% over the period analysed. By sector, most patents are concentrated in manufacturing, but emerging sectors – micro- and nanotechnologies, semiconductors, artificial intelligence and environmental technologies – show the highest rates of change, suggesting priority areas for digital and green transition policies. Patenting firms have average labour productivity of around €116,000 per employee, almost three times that of non-patenting firms, indicating a strong correlation between codified inventive activity and economic performance.

Andresciani's article (*From aggregate data to microdata. Private healthcare expenditure and tax incentives in Italy*) shows the challenges that arise when access to an administrative microdata archive

is not possible and a plurality of information sources must be used, each with its own strengths and limitations. The author's objective is to provide a critical assessment of the main challenges involved in measuring private healthcare expenditure, voluntary insurance coverage and tax deductions in Italy, highlighting how fragmentation of information, heterogeneity across sources and, in some cases, the high level of data aggregation may hinder the ability to monitor the equity and effectiveness of healthcare financing policies. The databases examined include national accounts for consumption classified by purpose (COICOP – Classification of Individual Consumption by Purpose), the Health Accounts System, the Istat Household Consumption Survey and administrative data from the Health Card System (STS). The national accounts and the Health Accounts System provide aggregate estimates of private expenditure, but differ in accounting framework, reference scope and treatment of co-payments or tax incentives, making comparisons across sources complex. The Istat Household Consumption Survey provides national and regional data, but is subject to sampling variability and possible under-reporting. The STS, managed by the Ministry of Economy and Finance, records healthcare transactions at individual level, but current publications provide only aggregated information, without distinguishing out-of-pocket (OOP) expenditure, insurance reimbursements or co-payments, thereby limiting distributional analysis. Unsurprisingly, the main findings therefore show that (i) estimates of private healthcare expenditure vary substantially depending on the source; (ii) the lack of microdata makes it impossible to distinguish who bears the expenditure, to what extent and with what insurance coverage, hampering analyses of territorial and socioeconomic equity; (iii) tax deductions are reported in aggregate form, making it impossible to identify whether the benefits relate to NHS co-payments or wholly private expenditure, thereby compromising assessment of their progressivity. Andresciani points out that, if the STS were made accessible at micro level for research purposes with appropriate privacy safeguards, it would be possible to break down private expenditure into OOP payments, insurance reimbursements and co-payments, and to assess the distribution of tax incentives across income groups and regions. Strengthening governance of healthcare expenditure data and access to more granular datasets therefore appear to be indispensable prerequisites for effectively monitoring the role of private financing in hybrid healthcare systems, assessing the actual “decongestion effect” of the private healthcare system, and ensuring that tax policies do not generate regressive effects.

The final two contributions to this special issue of Sinappsi illustrate the potential of administrative data sources also for so-called *ex ante* analysis. The article by **Bosia, Carlone, Floredan and Prever** (*Profiling people available for work in Piedmont. The role of administrative data and local territorial data*) assesses whether integrating individual administrative data with territorial information from Local Labour Market Systems (SLLs) can improve the predictive capacity of profiling models used by employment services, with particular reference to the Piedmont case. The administrative data used come from the Personal and Professional Record (SAP), a digital file that aggregates personal, educational, employment and participation-in-active-labour-market-policy information, automatically populated by Mandatory Communications and public employment services. The SAP covers the entire population of people who submit the Declaration of Immediate Availability (DID), ensuring complete coverage, high granularity and continuous updating. To enrich the analysis, the authors draw on the SLL information system produced by Istat, which combines business registers (ASIA), labour registers (including the Basic Individuals Register, RBI) and population registers to construct territorial indicators. These variables provide a detailed description of the productive and demographic structure of local areas. The main results show that the model replicated at regional level reproduces many of the regularities identified by the national model, but with marked differences: recent work experience has a stronger preventive effect on non-employment in Piedmont, while age and gender are less important determinants. Including SLL-level variables reduces prediction error, showing that a context with a high share of foreign workers increases the risk of failure to re-enter employment, whereas a higher presence of self-employed workers is associated with lower risk, provided that self-employment is structural. The results also suggest that the use of more flexible algorithms may identify people at higher risk of prolonged labour-market exclusion more effectively.

We thus come to the final contribution included in this special issue, by **Bazzoli, Marzadro, Tassi, and Fiorio** (*TREMODO-Imprese: a microsimulation model*), whose objective is to present TREMOD-Imprese,

a fiscal microsimulation model that makes it possible to assess ex ante business taxation policies at the level of the Autonomous Province of Trento. TREMOD-Imprese is based on a set of administrative data comprising IRES, IRAP and VAT tax returns for fiscal year 2019, integrated with the Istat register of active enterprises (ASIA). The tax data provide detailed information on profits, production value, turnover, tax rates, deductions and tax credits for each firm, while the business register provides size (employee classes), sector, registered office and other structural characteristics. Integration is performed using the firm's tax identification number, ensuring a one-to-one match between the three tax returns and the ASIA register. Access to these databases is made possible by the Tremod implementing agreement between the Statistical Institute of the Autonomous Province of Trento (ISPAT) and the Bruno Kessler Foundation, under the Memorandum of Understanding for the implementation of statistical research, which ensures information security and confidentiality. The illustrative simulations show that removing IRES deductions would increase revenue by €1.6 million (approximately <1% of IRES revenue), while eliminating reduced tax rates would increase it by €5.2 million ($\approx 2\%$). For IRAP, raising the standard rate to the national level (from 2.68% to 3.9%) would generate an increase of around €33 million ($\approx 33\%$ of IRAP revenue), whereas eliminating regional deductions would produce only a marginal increase of €1.1 million. The simulations are static, do not incorporate firms' behavioural responses, and are constrained by the lack of data on specific tax credits and on some variables needed for more complex scenarios. The authors conclude that integrating administrative tax data provides a useful tool for ex ante evaluation of tax policies also at sub-national level, but they stress the need to extend the time coverage and enrich the set of variables in order to enable more sophisticated and reliable simulations, for example by incorporating behavioural responses.

Open questions

The contributions collected in this issue of *Sinappsi* clearly demonstrate the potential of administrative data for producing empirical evidence and improving public policies, but they also highlight several obstacles that still limit their full exploitation in Italy. The open questions concern not only the physical availability of data, but also the legal and organisational conditions governing access, the ability to integrate different sources, the quality and documentation of archives, and the infrastructures needed to make them genuinely usable.

1. **Governance, coordination and common rules.** Authors highlight, first of all, the fragmentation of information systems and the absence, in some areas, of coordinated governance. The experience with labour-market data described by Anastasia, for example, shows differences in the quality of regional infrastructures, a lack of common standards and difficulties in building stable databases. Likewise, information sharing may be hindered by institutions' reluctance to make their archives available. According to the authors of the contributions published in this issue, fully exploiting administrative data therefore requires transparent access rules and coordination capable of fostering data sharing.
2. **Personal data protection and research access.** A second obstacle concerns the need to reconcile the use of data for statistical and research purposes with the protection of personal information. The regulatory framework described by di Somma allows data to be reused, but subjects such reuse to conditions of lawfulness, safeguards and limitations, while access to microdata requires specific procedures and protection measures, such as pseudonymisation, anonymisation, secure environments and monitoring of access practices. The open question is therefore how to make data accessible to research, within the framework of personal data protection rules, in a more systematic and transparent way than is currently the case. This is possible, as demonstrated by the numerous contributions submitted for this special issue and by other experiences in Italy and in other European Union countries subject to the same institutional constraints in force in Italy (Berton and Paruolo 2026).
3. **Interoperability, linkage and standardisation.** The fragmentation of databases makes the ability to link information from different sources central. The contributions devoted to interoperability and record linkage nevertheless show that this process is not automatic: sources may not share unique identifiers (Priulla *et al.*; Salamone) and may differ in structure, definitions and informational

characteristics. Buonocore *et al.* therefore stress the importance of semantic standardisation, controlled vocabularies and data schemas, as well as more stable linkage infrastructures.

4. **Data quality, documentation and granularity.** A further problem concerns the availability of sufficiently detailed and well-documented information, namely microdata at the level of the individual statistical unit. Andreschiani's contribution on private healthcare expenditure highlights in particular the limitations arising from access only to aggregate data: the absence of microdata makes it impossible to distinguish who bears the expenditure, to what extent and with what forms of coverage, hampering distributional analysis. Other contributions instead point to the need to improve dataset documentation, data quality and the continuity of published information. Data availability is therefore not sufficient if its granularity, quality or documentation does not allow for adequate analytical use.
5. **Infrastructure, resources and skills.** The full development of statistical analysis also requires sustained investment in physical capital (secure infrastructures) and human capital (technical and organisational resources within data-holding institutions, as well as training opportunities for researchers). Even the most structured experience in the Italian context, namely VisitINPS (De Paola *et al.*), identifies broader access, greater interoperability, wider dissemination of anonymised or synthetic microdata and stronger documentation as areas for development. The mere existence of administrative archives is therefore not sufficient; the technical and professional capacity needed to turn them into a resource that can actually be used by researchers must also be in place.
6. **Continuity and integration of data into policy processes.** Finally, authors stress that data availability must be accompanied by continuity in publication and by a structured connection with decision-making cycles. The contribution by Bosia *et al.* shows how integrating administrative data with territorial data can improve the decision-making processes that currently guide policy delivery. The OpenCoesione experience (Marinuzzi and Tortorella), on the other hand, shows that simply publishing datasets does not automatically generate impact: data must be comparable, actionable and available with regular updates. In turn, this depends on a mix of obligations and incentives for using administrative data for monitoring and evaluation that involves all parties concerned: policymakers, holders of raw data and users of data for statistical purposes.

We conclude this introduction to the first part of the contributions devoted to the use of and access to administrative data sources for research purposes in Italy by asking why, despite the advantages offered by such sources and described above, Italy still finds itself, broadly speaking, only halfway there, with many interesting experiences from an analytical perspective and a fairly well-established tradition, but with an overall system that remains too fragmented and insufficiently structured to exploit its full potential. We do not believe that the reasons lie in regulations protecting personal and sensitive data. Some experiences in Italy and elsewhere within the European Union show that, although difficult, building a system for using administrative data sources for research purposes that combines compliance with the rules and flexibility in access arrangements is far from impossible. Nor do we believe that this shortcoming can be attributed to what Berton and Paruolo (2026) call "false beliefs" about the use of administrative data for research. First, as regards the cost of preparing such sources for research purposes: while it is certainly true that, unlike data originating from surveys, administrative sources require an initial investment in cleaning, documenting and preparing raw data, this cost is largely fixed and does not have to be incurred again for every new data release, unlike surveys, whose cost structure is dominated by the variable component. Second, as regards delays in the availability of data that can be processed for analytical purposes, which are supposedly incompatible with the constraints of policymaking, we believe the opposite is true: the generation of raw administrative data is continuous by construction, and transforming them into information usable for analysis requires, once the relevant procedures have been validated, at most a few weeks; by contrast, the validation and fieldwork timelines for questionnaires are at least several months and generally follow a predetermined schedule that is difficult to adapt to emerging policy information needs.

The factors that are readily identified in the current debate as the 'usual suspects' – privacy rules, costs, and technical or time constraints – can therefore explain only part of the delays currently

observed in Italy. Deeper reasons may be found by extending the concept of 'data jealousy' introduced by Anastasia to refer to a more general lack of willingness and an inadequate incentive system to encourage the authorities responsible to invest in a coordinated and structured system for research access to administrative data sources, with clear and uniform rules and access arrangements whose costs are sustainable for data users.

References

- Berton F., Paruolo P. (a cura di) (2026), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer (forthcoming)
- Bloem M., Gorissen I., van den Boom R., van Hees V. (2026), CBS microdata: a user perspective, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 21 (forthcoming)
- Commissione europea (2026), *Review of the 2021 European Pillar of Social Rights Action Plan*, Bruxelles, SWD (2026) 650 final
- Consiglio dell'Unione europea (2024), *Draft Council Conclusions on the role of labour market, skills and social policies for resilient economies*, Bruxelles, 11066/24
- Contini B., Trivellato U. (a cura di) (2005), *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*, Bologna, il Mulino
- Crato N., Paruolo P. (a cura di) (2019), *Data-driven policy impact evaluation. How to access microdata is transforming policy design*, Berlino, Springer
- Fabre B., Gadouche K. (2026), Access to French administrative microdata, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 11 (forthcoming)
- Leombruni R., Ardito C. (2026), The long and winding road. The access to administrative data for research in Italy, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 16 (forthcoming)
- Lima F., Martins J.P., Neugebauer K., Pimentel J.F., Rodrigues S., Saraiva P., Sousa R., Zilhão M.J. (2026), Administrative data in Portugal, in Berton F., Paruolo P. (eds.), *Data-driven learning in the EU. From administrative sources to public value*, Berlin, Springer, Chapter 23 (forthcoming)