INAPP
PUBLIC POLICY INNOVATION

23 giugno 2017

Summary

I seminari del venerdì

Il trattamento degli archivi amministrativi per la ricerca: l'esperienza del progetto Visit_INPS Scholars 2015-2016 sulle prestazioni di sostegno al reddito

Autore

Roberto De Vincenzi

INTRODUZIONE

Il seminario illustra i risultati raggiunti dal progetto realizzato all'interno del programma Visit INPS Scholars (edizione 2015 - 2016) dal titolo: "Interoperabilità delle banche dati amministrative sulle politiche passive e attive del lavoro". Coerentemente con l'impostazione originaria, la realizzazione del progetto ha collocato al centro della propria azione la costruzione delle condizioni indispensabili alla integrazione di differenti archivi amministrativi.

L'interoperabilità, dunque, è intesa come obiettivo di medio lungo periodo, come punto d'arrivo dell'indispensabile percorso metodologico, tecnico, ma anche d'interpretazione semantica, che sinteticamente viene definito "trattamento e normalizzazione" di un archivio amministrativo.[1]

In termini più specifici, la finalità operativa del progetto e il senso dei risultati raggiunti risiede nel contributo diretto alla predisposizione, a partire dalla fornitura originaria (l'intero archivio di microdati grezzi in formato csv), di un database direttamente utilizzabile a fini statistici e conoscitivi sia da INPS che da altre istituzioni, accademiche e delle ricerca pubblica.[2]

In questo specifico caso il percorso conoscitivo e di normalizzazione è stato realizzato sull'archivio delle politiche passive del lavoro, sul database dell'INPS denominato Banca dati Percettori, che archivia le informazioni circa gli individui e le aziende di riferimento per ciascun trattamento di sostegno al reddito offerto ai lavoratori licenziati o sospesi temporaneamente. Si tratta quella porzione di politiche pubbliche per il lavoro che per dimensione economica e politica è al centro degli interessi conoscitivi della collettività.

La Banca dati Percettori dell'INPS è un archivio di microdati sulle prestazioni di sostegno al reddito strutturata, in termini di unità elementare, per singolo trattamento. In tale archivio, entrato a regime nel 2009, confluiscono tutte le informazioni anagrafiche concernenti l'individuo trattato, le informazioni riguardanti il trattamento (tipologia, date e importi) e le informazioni concernenti le aziende di riferimento registrate nei diversi sistemi informativi gestionali utilizzati dagli operatori INPS dislocati presso le 3000 sedi dell'Istituto di previdenza su tutto il territorio nazionale.

Benché più numerosi, i sistemi gestionali utilizzati dall'INPS afferiscono a due gruppi principali: i sistemi di gestione amministrativa delle prestazioni di sostegno al reddito in caso di disoccupazione (attualmente le principali forme sono NASPI, DIS_COLL e la Mobilità) e le prestazioni di sostegno al reddito in caso di sospensione temporanea dal lavoro (la CIG). Si tratta di prestazioni molto differenti tra loro per le procedure amministrative, per i target di popolazione raggiunta e per il ruolo svolto dalle imprese. Tale differenza si ripercuote sui connotati e sulla struttura dei flussi informativi che alimentano l'archivio generale che li contiene entrambi. Lo sforzo realizzato dall'INPS è stato quello di creare un unico e razionale contenitore (la Banca dati Percettori) di informazioni , le quali, proprio in base alle due

[1] Cfr. G. Brancato (a cura di), 2016, ISTAT, *Qualità dei processi statistici che utilizzano dati amministrativi*; Bernardi, Cerroni, De Giorgi – 2013 ,ISTAT -DARCAP – *Documentazione degli Archivi delle PPAA*; M.Calzaroni , 2015, ISTAT (Progetto SIM -Sistema Integrato Microdati - 2015); B.Anastasia - Osservatorio e Ricerca Veneto Lavoro (SILV e Progetto Planet 2.2 – 2015.

[2] L'Istituto previdenziale, titolare dell'archivio , inoltre, è destinatario di specifico materiale tecnico (procedure e script) che intende documentare il percorso conoscitivo (metodologico, tecnico e semantico) affrontato nella realizzazione del progetto.

macro tipologie di prestazioni (in caso di disoccupazione e in caso di sospensione temporanea dal lavoro) spesso assumono significati differenti.[3]

1 IL PROCESSO DI NORMALIZZAZIONE DELLA DEL DATA BASE

La fornitura originaria, composta da oltre 17 milioni di record è stata importata all'interno del software DBMS[4] messo a disposizione dall'INPS ed è stata trattata secondo le procedure di normalizzazione necessarie alla individuazione e gestione dei record:

- incompleti e non validi ;
- duplicati (duplicati "fotocopia" e duplicati incoerenti)

Record inconpleti/non validi. Per la gestione dei record incompleti e non validi si è proceduto alla loro individuazione attraverso semplici interrogazioni sui campi ritenuti indispensabili al loro successivo trattamento (ID_soggetto, tipo di prestazione, data di decorrenza, importo e durata in ore per la CIG) o in giorni per li sussidi in caso di disoccupazione). L'assenza di queste informazioni ha comportato l'eliminazione di tali record dal data base e la contestuale archiviazione in tabelle separate.
Inoltre, nello stesso data base di microdati grezzi, è stata utilizzata l'informazione gestionale "prestazione cancellata" utilizzata dagli operatori INPS per memorizzare (temporaneamente o definitivamente) record non validi. In sintesi, rispetto al totale dei record forniti originariamente sono stati individuati circa 470 mila record incompleti/non validi, pari al 2,7%.

Record duplicati: la procedura utilizzata per la individuazione e gestione dei record duplicati è stata impostata a partire dalla definizione di incoerenza che deriva dalle norme e dalle derivanti procedure amministrative: "ogni individuo (lavoratore) può beneficiare, nello stesso periodo (data decorrenza trattamento e data fine), esclusivamente di un solo trattamento". Questo assunto, rispetto alle tipologie di trattamento presenti in Banca dati Percettori, è sempre vero fatta eccezione per i trattamenti (300 mila trattamenti per altro inseriti di recente) concernenti il *Fondo integrativo del Trasporto aereo* e il cosiddetto *Assegno emergenziale Fondo credito*[5]. A parte la presenza di un numero esiguo di duplicati identici o fotocopia, la metà dei quali (922 record) eliminati dalla Banca dati - sono stati individuati circa 486 mila record duplicati incoerenti ossia con almeno lo stesso ID_soggetto e la stessa DT_decorrenza prestazione. Il passo successivo alla loro individuazione (basata sull'assunto sopra esplicitato) è stato quello definire un criterio (il più verosimile possibile) di selezione dei record "veri o buoni" dai record "falsi o non buoni".

[3] L'esempio più intuitivo è dato dai diversi campi dedicati alle informazioni sull'azienda di riferimento che per i trattamenti di disoccupazione informano circa l'ultima azienda presso il quale ha lavorato il lavoratore disoccupato trattato, mentre per la CIG gli stessi campi informano circa il datore di lavoro del lavoratore sospeso temporaneamente.
[4] I software DBMS (*Database Management System)* sono sistemi dedicati alla creazione, manipolazione e interrogazione efficiente di database (ovvero di collezioni di dati strutturati). Su richiesta degli autori, l'INPS ha messo a disposizione il software open source denominato MYSQL Server.
[5] I record di quest'ultimo tipo (che abbiamo denominato "duplicati coerenti") sono stati comunque salvati su una tabella separata perché avrebbero compromesso l'efficienza delle procedure successive.

La presenza, nella Banca dati percettori di campi gestionali (cioè che registrano, spesso in modo automatico, informazioni utili all'operatore INPS che gestisce la prestazione ha permesso di individuare un criterio di scelta dei record "buoni" da quelli "non buoni". Ci riferiamo in particolare alla data di ultima modifica di ciascun record (campo TS_Variazione) e all'assunto, fatto nostro dopo esser stato condiviso con i referenti interni all'INPS, che il record da mantenere tra due duplicati sia quello più aggiornato[6], ovvero quello con la data più recente. Ciò ha reso possibile scegliere i 256.215 duplicati incoerenti (pari all'1,5% dei record totali presenti nella fornitura originaria) eliminati dalla Banca dati percettori e salvati su una tabella a parte.

In ultima analisi, il processo di normalizzazione ha individuato nel complesso 727.490 record incompleti o duplicati (pari al 4,2% del totale dei record originari) eliminati dal data base e salvati su tabelle separate. Stesso trattamento è stato riservato ai 380.113 record riguardanti i Fondi integrativi di cui sopra.

2 CONTENUTI INFORMATIVI E LIMITI STRUTTURALI ATTUALI DELL'ARCHIVIO

Alla data di redazione del presente report la Banca dati Percettori dell'INPS, attraverso un aggiornamento costante, archivia le informazioni su trattamenti, percettori e aziende di tutte le prestazioni di sostegno al reddito in caso di disoccupazione, alle quali si aggiungono le informazioni su percettori, trattamenti e aziende di alcune speciali tipologie di trattamento cofinanziate da risorse comunitarie (programma Garanzia Giovani, Programma Pari e Programma Welfare to Work).

Per quanto concerne invece i trattamenti di sostegno al reddito in caso di sospensione temporaneo, nella Banca dati Percettori risultano registrate esclusivamente i trattamenti di Cassa integrazione Guadagni Straordinaria (CIGS) a pagamenti diretti e la Cassa integrazione Guadagni in deroga (GIGD). In definitiva per la CIG in genere sono presenti tutte le prestazioni economiche erogate direttamente dall'INPS ai lavoratori dipendenti sospesi temporaneamente.

Nell'archivio INPS, per il momento, sono del tutto assenti i trattamenti di CIG cosiddetta a conguaglio (ossia le prestazioni per sospensione anticipate dai datori di lavoro successivamente conguagliate degli importi erogati dall'INPS). I trattamenti mancanti[7] coincidono con la totalità delle prestazioni di Cassa integrazione Guadagni Ordinaria (CIGO) e una quota (si stima circa la metà o poco più) delle prestazioni di CIGS, la quota parte delle prestazioni erogate a conguaglio (anticipate dall'azienda successivamente rimborsata).

In sintesi, alla data di estrazione del data base da parte dell'INPS (fornitura dei dati grezzi) la Banca dati Percettori contiene i microdati relativi all'universo dei trattamenti di sostegno al reddito (relativamente agli anni 2009-2016) in caso di disoccupazione involontaria (o nei casi previsti dalle norme anche volontaria) e una sottopopolazione consistente (identificabile con le aziende con maggiori

[6] Il campo TS_variazione è un campo propriamente gestionale che registra automaticamente la data e l'ora ogni qualvolta l'operatore INPS modifica e salva il record in lavorazione.

[7] Tale limite, secondo le previsioni della DC Organizzazione e Sistemi informativi – area Ammortizzatori Sociali dell'INPS, sarà a breve superato grazie all'entrata a regime della nuova procedura di richiesta di autorizzazione per il ricorso alla Cassa integrazione guadagni da parte delle imprese prevista introdotta dal D.Lgs. 148/2015.

difficoltà economiche) di trattamenti di sostegno al reddito in costanza di rapporto di lavoro (CIG a pagamenti diretti).

La fornitura originale (microdati grezzi) è stata depurata dai record incompleti e duplicati fino alla definizione del database normalizzato - denominato SIP_DATA - **composto da 16.250.818 record validi** (si ricorda che ciascun record equivale ad un trattamento). Di questi, circa 800mila trattamenti sono precedenti al 2009 e poco più di 172mila trattamenti sono invece successivi al dicembre 2016. I restanti record si distribuiscono lungo l'arco temporale che va gennaio 2009 al dicembre 2016, identificabile come il periodo rispetto al quale la Banca dati Percettori, previo trattamento di normalizzazione, offre dati stabilizzati e statisticamente trattabili.

Circa il contenuto informativo di dettaglio della Banca dati Percettori si si illustrano sinteticamente i seguenti elementi:

- l'archivio, prima di essere fornito, è stato preventivamente anonimizzato dall'INPS. Prima di essere anonimizzati, dai Codici fiscali degli individui sono state estratte le informazioni anagrafiche in esso contenute (genere, luogo e data di nascita) e memorizzate su campi dedicati;
- le date di carattere gestionale (DT) registrate nella Banca dati percettori rappresentano un'informazione quanto mai strategica utilizzate per molteplici obiettivi (normalizzazione, controllo, stime, individuazione sottopopolazioni, etc.) Ci si riferisce alla già citata TS_variazione, ma anche alla DT_cessazione (del contratto i lavoro), la DT_decorrenza (della prestazione), la DT_fine teorica (della prestazione) e la DT_fine_effettiva (sempre della prestazione);
- come in ogni archivio amministrativo complesso le informazioni sono strutturate e collegate tra loro secondo un modello relazionale che è stato parzialmente recuperato dal trattamento attraverso la produzione di una tabella principale (denominata SIP_DATA) e di tabelle secondarie quali l'archivio delle aziende (denominato SIP_SEDI_DL) a vario titolo coinvolte nei trattamenti di sostegno al reddito e le tabelle di metadati (settori, comuni, provincie, etc);
- la Banca dati Percettori è un oggetto relativamente dinamico. Nel corso degli ultimi 5 anni (sulla base di quanto è stato possibile appurare per una sottopopolazione di trattamenti)[8] l'archivio si è progressivamente arricchito di informazioni, spesso in connessione con le specificità gestionali delle tipologie di prestazione introdotte dalla normativa;
- dal set di informazioni originarie sono state prodotte ed inserite nel data base principale normalizzato (SIP_DATA) alcune nuove variabili (prodotte sulla base delle informazioni originarie) valutate come indispensabili ad un più efficiente utilizzo statistico delle informazioni;
- è stata prodotta una tabella di storie individuali intra-Banca dati Percettori (denominata STORIE_PERCETTORI) che per ciascun individuo ricostruisce, su base mensile, tutti i trattamenti (a prescindere dal tipo di trattamento) occorsi dal 2009 al 2016 nonché i giorni di attesa tra la fine di un trattamento e l'inizio eventuale di quello successivo e le informazioni sull'azienda di riferimento.

8 Cfr 1. De Vincenzi R., Irano A. e Sorcioni M. (a cura di), Ammortizzatori sociali in deroga e politiche attive del lavoro: l'attuazione, gli esiti e gli effetti dell'Accordo Stato Regioni 2009-2012, I libri del FSE, Voll. I-II, Rubettino, Soveria Mannelli, 2014.

3 ANALISI DESCRITTIVE

Il report conclusivo dello svolgimento del progetto la cui consegna alla Direzione Centrakle Studi e ricerche dell'INPS è prevista per il 28 giugno p.v., e che sarà successivamente pubblicato sulla collana «WorK INPS Progress», dedica una parte alla descrizione dei dati presenti nella Banca dati Percettori.
Le analisi descrittive, per informazioni anagrafiche, per caratteristiche dell'ultimo lavoro svolto, per durate e consistenza delle prestazioni sono state concentrate sulle prestazioni di sostegno al reddito in caso di disoccupazione; quelle quantificate al solo fine di offrire un dato sintetico nella tabella successiva.

Tabella - Trattamenti per tipo di prestazione e anno di cessazione e totale percettori per anno di cessazione

Trattamenti per anno - DT_CESSAZIONE anno	2009	2010	2011	2012	2013	2014	2015	2016
Indennità di mobilità ordinaria	58.778	62.460	58.182	64.451	81.158	126.670	50.547	35.211
Indennità di mobilità in deroga	19.035	33.442	41.073	51.726	33.832	11.377	409	1.402
Disoccupazione ordinaria con requisiti normali	884.981	911.179	981.418	1.124.669				
Disoccupazione ordina e trattamento speciale edilizia L 427/1997	31.699	14.293	11.322	9.133				
Disoccupazione lavoratori marittimi	1.866	1.294	487	194				
Disoccupazione sospesi	46.151	15.226	25.003	66.089				
ASpI					1.076.058	1.071.850	260.238	
Mini ASpI					449.229	545.231	107.476	
ASpI sospesi					22.548	24.265	20.268	
NASPI							1.300.739	1.519.479
DISCOLL							18.344	8.646
TOTALE TRATTAMENTI	**1.042.510**	**1.037.894**	**1.117.485**	**1.316.262**	**1.662.825**	**1.779.393**	**1.758.021**	**1.564.738**
TOTALE PERCETTORI	**975.828**	**964.680**	**1.037.547**	**1.205.015**	**1.599.399**	**1.698.064**	**1.735.576**	**1.519.856**

Fonte: elaborazioni proprie su microdati Banca dati Percettori dell'INPS

L'attenzione è stata poi focalizzata sui trattamenti di NASPI (Nuova Assicurazione Sociale per l'Impiego) ossia sul più importante strumento di sostegno al reddito in caso di disoccupazione attualmente in vigore. Le analisi descrittive di profondità hanno messo il luce una serie di elementi conoscitivi indispensabili all'interpretazione dei dati e all'uso corretto di metodologie e tecniche di analisi di tipo avanzato.
Ci si riferisce in particolare a:

- la necessità di **evitare l'utilizzo dei microdati amministrativi (in genere e nella fattispecie) senza aver verificato le distribuzioni (semplici) descrittive e ricercato il significato di eventuali anomalie**. Solo per fare due esempi, abbiamo potuto appurare che la mancanza di informazioni sull'azienda presso il quale ha lavorato l'individuo trattato dalla NASPI è nella quasi totalità dei casi concernente i trattamenti rivolti a *badanti e colf* (156 mila trattamenti, nel 91% dei casi donne, di età media di poco superiore ai 50 anni) O ancora, la peculiarità delle informazioni inserite nelle domande di sussidio NASPI presentate dai *docenti precari della scuola* (specie *secondaria superiore,* pari a 300mila trattamenti, il numero più altro di trattamenti per gruppi Ateco ISTAT). In quest'ultimo caso piuttosto emblematico, solo un approfondimento

d'indagine di tipo qualitativo (svolto sui blog della scuola) ha potuto accertare che, nella gran parte di questi casi, nelle domande di sussidio è stato indicato come datore di lavoro (con il Codice fiscale) il CED del Ministero dell'Economia che ha sede a Latina il quale, in base ad una convenzione sottoscritta con il MIUR, da diversi anni predispone le buste paga dei docenti sella scuola pubblica che da lì hanno rilevato (erroneamente) e trascritto nella domanda il codice identificativo del datore di lavoro.[9]

- la necessità di **un'analisi specifica di alcune dimensioni in gran parte inedite per le precedenti misure di sostegno al reddito** come, ad esempio, la sospensione temporanea della prestazione (per l'avvio di un contratto di lavoro a termine dalla durata e dalla retribuzione relativamente contenuta), oppure l'interruzione della prestazione per la perdita dei requisiti (avvio di un lavoro più consistente in termini di durata e retribuzione).

Tabella - Trattamenti NASPI dal 9 maggio 2015 (prima data decorrenza registrata) al 15 marzo 2017 (ultima decorrenza registrata

	Totale trattamenti	Di cui conclusi	%
Totale trattamenti	**2.884.343**	**1.520.582**	**52,7%**
Nell'annualità 2015	1.185.151	869.364	73,4%
Nell'annualità 2016	1.552.522	644.296	41,5%
Primi tre mesi 2017	146.670	6.922	4,7%
Femmine	1.499.405	806.976	53,8%
Maschi	1.384.938	713.606	51,5%
Trattamenti conclusi a scadenza naturale		**746.420**	**dt_fine effettiva = dt_fine teorica**
Trattamenti conclusi che hanno beneficiato di sospensione che non si è trasformata in interruzione		**131.682**	**dt_fine effettiva > dt_fine teorica**
Trattamenti interrotti per perdita requisiti (lavoro stabile o consistente)		**642.480**	**dt_fine effettiva < dt fine teorica**

Fonte: elaborazioni proprie su microdati Banca dati Percettori dell'INPS

- Tale analisi tende a orientare gli **interessi verso la definizione di profili di disoccupazione maggiormente ancorati ai reali "episodi lavorativi"** la cui informazione puntuale, insieme ad altre informazioni strategiche (come il livello d'istruzione o il reddito familiare o la partecipazione alle politiche attive), andrebbe integrata a questo importante archivio, le cui informazioni comunque offrono nuove e interessanti consapevolezze.

[9] Alla richiesta di chiarimenti espressi sui blog dei docenti precari della scuola pubblica dai richiedenti NASPI circa l'identificazione della matricola del datore di lavoro, l'indicazione, spesso proveniente da esperti e consulenti, ("la trovate sulla busta paga") è valida e corretta per tutti i lavoratori dipendenti ad esclusione (per la convenzione in essere tra CED del Ministero dell'Economia e MIUR) dei docenti della scuola pubblica. Quindi il problema informativo andrebbe risolto con una specificazione più puntuale (rivolta ai docenti precari) sul modulo della domanda di NASPI.

Tabella - NASPI concluse per tipo di conclusione (prima, nature o dopo fine teorica), classi di durata posticipo o anticipo e classi d'età dei trattati

Datediff (DT_fine_eff.iva, DT_fine_teorica)	Durata	Giovani fino a 29 anni	Età media 30-44	Età media avanzata 45-54	Età avanzata 55 - 60 e oltre	NULL	Totale	Totale per gruppi (dopo, nature e prima)	
Dopo fine naturale	Fino a 1 mese	3,7	3,3	3,2	2,5		3,3		
	Da 1 a 3 mesi	3,1	2,9	2,8	2,3		2,8		
	Da 3 a 6 mesi	2,4	2,1	1,7	1,5		2,0	**8,7**	131.682
	Da 6 a 12 mesi	0,6	0,6	0,4	0,3		0,5		
	Oltre 12 mesi	0,0	0,0	0,0	0,0		0,0		
	Cessate dopo totale	*9,7*	*8,8*	*8,2*	*6,7*				
Nature	Nature	60,1	44,5	46,4	49,4		49,1	**49,1**	746.420
Prima fine naturale	p_fino a mese	4,0	4,2	4,2	3,6		4,1		
	p_da 1 a 3 mesi	13,0	14,8	13,5	12,8		13,8		
	p_da 3 a 6 mesi	7,8	13,1	12,5	12,0		11,6	**42,3**	642.480
	p_da 6 a 12 mesi	3,3	8,5	8,7	8,8		7,4		
	p_oltre 12 mesi	2,1	6,1	6,6	6,8		5,4		
	Cessate prima totale	*30,2*	*46,7*	*45,4*	*43,9*				
	Totale complessivo	**100,0**	**100,0**	**100,0**	**100,0**		**100,0**	**100,0**	**1.520.582**
	Totale complessivo	**354.464**	**649.670**	**351.566**	**164.875**	**7**			

Fonte: elaborazioni proprie su microdati Banca dati Percettori dell'INPS

In conclusione, benché nel seminario sarà dedicato maggiore spazio alla illustrazione degli ulteriori approfondimenti descrittivi anche in riferimento alle "storie dei percettori" (intra Banca dati Percettori), appare opportuno rimandare al report conclusivo l'analisi puntuale del **percorso metodologico, tecnico e d'interpretazione semantica delle informazioni archiviate nel sistema informativo nazionale sulle politiche passive del lavoro** finalizzato alla costruzione delle condizioni indispensabili alla integrazione di differenti archivi amministrativi.

Il programma Visit_INPS Scholars e le connesse risorse messe a disposizione, nonché la costante disponibilità delle strutture tecniche dell'Istituto previdenziale hanno permesso a questo progetto di completare tale percorso (in modo reputato soddisfacente dalla stesso Istituto promotore)[10] su una porzione ampia ancorché complessa (quale è la Banca dati Percettori dell'INPS) del futuro sistema informativo unitario sulle politiche del lavoro.

[10] A tale riguardo, è intenzione della DC Studi e Ricerche dell'INPS fornire la Banca dati Percettori normalizzata (e non più i microdati grezzi) ai ricercatori interessati e vincitori della seconda edizione del Bando Visit_INPS (le cui proposte progettuali sono in corso di valutazione).