



VENETO LAVORO
Osservatorio & Ricerca

PUF 3.0 – GUIDA A MERCURIO
Storia, contenuto e specifiche

Gianluca Emireni

Aprile 2016

1. Storia del Puf (Public use file): da Giove a Planet, da Planet a Mercurio

Mercurio è il nuovo database statistico (la terza versione), costruito con i dati aggiornati a dicembre 2015, che l'Osservatorio di Veneto Lavoro mette a disposizione dei ricercatori – come Puf (**Public use file**) - per metterli in grado di sfruttare a pieno il bagaglio di informazioni derivanti dal sistema gestionale amministrativo dei Centri per l'impiego, sistema che organizza i flussi delle comunicazioni obbligatorie (CO) e delle dichiarazioni di disponibilità al lavoro rilasciate dai lavoratori. Nella sua prima versione il **Puf, denominato “Giove”**, ha rappresentato un'esperienza innovativa di accesso alle basi dati amministrative, tra le prime a livello nazionale a mettere a disposizione del mondo della ricerca informazioni a livello elementare nel rispetto della legge sulla privacy e del codice deontologico.

La principale fonte di alimentazione di “Giove” era il sistema **NetLabor**, in uso presso i Centri per l'Impiego del Veneto. Le caratteristiche di quel sistema informativo richiedevano una consistente lavorazione del dato, a causa di modalità di interrogazione e tracciati record condizionati dall'organizzazione “amministrativa” delle informazioni. In Giove non erano disponibili informazioni su e disoccupazione e liste di mobilità.

La prima release di Giove risale al 2004, l'ultima al **febbraio 2008**.

La svolta che ha apportato un cambiamento radicale nell'organizzazione delle informazioni che pervengono ai Centri per l'impiego in materia di rapporti di lavoro è stata segnata dall'avvento a livello nazionale del sistema di **Comunicazioni Obbligatorie (CO)** e, per il Veneto, dall'adozione del **Sistema Informativo del Lavoro (Silv Veneto)**.

Queste modifiche di contesto hanno portato l'Osservatorio di Veneto Lavoro a rivedere le modalità di costruzione del Puf, dando luogo alla **versione 2**, denominata **Planet**.

Rispetto a Giove, Planet ha comportato un affinamento delle modalità di elaborazione, correzione e distribuzione del dato statistico. Ha visto il rilascio di versioni successive, che hanno mantenuto la medesima struttura, pur aggiornandosi nei contenuti e riportando correzioni sempre più affidabili (e di assoluta trasparenza per il ricercatore-utilizzatore).

Le differenze con **Giove** sono numerose. Il tracciato record di Planet e le modalità di interrogazione sono completamente cambiate; inoltre la versione 2 inglobava mobilità e disoccupazione in coerenza con i rapporti di lavoro e mettendo a disposizione una tabella per seguire i cambiamenti di status di tutti i lavoratori presenti nella base dati.

Sono rimaste inalterate nel tempo le modalità di distribuzione, adottate anche da **Mercurio**: in forma anonimizzata o con i dati in chiaro per consentire un possibile link ad altre basi dati in possesso dei ricercatori-utilizzatori qualora ciò sia esplicitato nel progetto di ricerca alla base della richiesta del Puf.

Il rilascio del Puf, tanto più in funzione di quest'ultima modalità di diffusione, è subordinato alla presentazione, oltre che di un adeguato progetto di ricerca, della documentazione prevista dalla normativa in materia di protezione dei dati personali, il tutto sottoposto alla valutazione e approvazione da parte di una Commissione appositamente istituita presso Veneto Lavoro.

La prima release di Planet risale al maggio 2008, l'ultima al **luglio 2015**.

Il consolidamento del sistema Silv ed alcune modifiche introdotte nelle CO, assieme alla volontà di semplificare l'interrogazione del database statistico, hanno portato l'Osservatorio a realizzare la terza versione del Puf, denominata **Mercurio**¹.

La prima release del Puf 3.0 è di aprile del 2016, con dati aggiornati a fine 2015.

¹ Nella mitologia greca Mercurio (Hermes), figlio di Zeus e della ninfa Maia, era il **messaggero degli dei, dio protettore dei viaggi e dei viaggiatori, della comunicazione** [...].

2. L'organizzazione generale di Mercurio

Il nuovo Puf condivide con Planet la fonte dell'informazione (Silv e CO) ma prevede una diversa organizzazione nelle tabelle dei rapporti di lavoro, disoccupazione e mobilità.

I rapporti di lavoro

Nelle precedenti versioni i rapporti di lavoro erano organizzati in diverse tabelle, ognuna delle quali descriveva una particolare tipologia di evento: assunzioni, cessazioni, trasformazioni e proroghe. Questo garantiva una descrizione dettagliata dell'evoluzione del rapporto di lavoro ma rendeva laboriose le operazioni di recupero e allineamento dell'intero flusso di accadimenti (un rapporto di lavoro prevede necessariamente un evento di assunzione, può includere una trasformazione e/o una o più proroghe ed eventualmente una cessazione).

In Mercurio ogni rapporto di lavoro è descritto all'interno di un singolo record in un'unica tabella (quella appunto dei rapporti di lavoro). Ciò permette di mantenere inalterato il contenuto informativo necessario a ricostruire la storia del rapporto e facilita il recupero delle informazioni.

La disoccupazione

Planet descriveva gli episodi di disoccupazione solamente tramite le **dichiarazioni di disponibilità (Did)** dei lavoratori. Queste però non risultano sufficienti a distinguere i periodi di disoccupazione "amministrativa" da quelli di effettivo "non lavoro" (episodi distinti di disoccupazione).

Mercurio analizza le dinamiche di passaggio tra stati di occupazione e disoccupazione, include il motivo di ingresso e di uscita dalla disoccupazione ed esclude i periodi di lavoro in regime di disoccupazione amministrativa (particolarmente rilevanti in passato).

Alla luce del frequente cambiamento delle regole degli stati di disoccupazione probabilmente questo dataset sarà quello che più spesso verrà modificato per far fronte alle novità normative.

La mobilità

I flussi di mobilità sono gestiti in modo del tutto analogo rispetto alla versione precedente.

3. Mercurio in cifre

Con riferimento alla prima versione di Mercurio, il volume di dati disponibile è il seguente:

- **15 milioni** di rapporti di lavoro, relativi a **3,3 milioni** di lavoratori e **800 mila** datori di lavoro;
- **3,5 milioni** di flussi di entrata in disoccupazione, relativi a **1,8 milioni** di Did (Dichiarazioni di disponibilità) e a **1,25 milioni** di lavoratori coinvolti;
- **360 mila** domande di mobilità, inoltrate da **322 mila** lavoratori.

	Fine										
	Fino al 2000	2001-2005	2006-2009	2010	2011	2012	2013	2014	2015	Rapporto aperto	Totale
Fino al 2000	2.360.419	481.476	148.627	30.112	25.708	23.558	18.890	22.015	22.513	190.194	3.323.512
2001-2005		2.248.305	389.730	31.066	24.974	21.446	17.808	16.308	15.490	200.759	2.965.886
2006-2009			2.622.107	256.810	91.023	57.334	40.647	31.123	26.625	293.297	3.418.966
Inizio 2010				551.450	184.568	38.179	20.866	12.691	9.966	68.906	886.626
2011					591.626	183.042	36.596	20.170	14.394	81.767	927.595
2012						578.544	166.715	33.968	20.840	82.815	882.882
2013							559.322	162.750	34.140	86.703	842.915
2014								602.456	169.670	120.254	892.380
2015									631.853	338.999	970.852
Totale	2.360.419	2.729.781	3.160.464	869.438	917.899	902.103	860.844	901.481	945.491	1.463.694	15.111.614

4. Data quality sul rapporto di lavoro

I gestionali di Silv e CO - le due fonti di alimentazione di **Mercurio** – sono oggetto di continue modificazioni e correzioni, sia dal punto di vista logico, sia per quanto riguarda il sistema di validazione (oltre che per tenere conto delle novità normative).

Nonostante alcuni dei problemi riscontrati negli anni scorsi siano stati risolti in fase di validazione delle Comunicazioni Obbligatorie, permangono parziali incongruenze, determinate soprattutto

- dalla definizione di algoritmi deterministici, a volte un po' rigidi, rispondenti all'esigenza di ridurre al minimo la quantità di comunicazioni “da validare manualmente”,
- dalla bassa qualità delle informazioni ereditate dal precedente gestionale (Netlabor).

In questa versione del Puf abbiamo sviluppato degli algoritmi in grado di correggere, dove possibile, situazioni di palese incertezza, cercando – nei casi dubbi - di mantenere un atteggiamento “conservativo” per evitare di modificare in modo eccessivo l'informazione originale.

Le modifiche riguardano diversi aspetti e possono introdurre delle modifiche nei dati o, nel caso in cui risulta impossibile ricostruire correttamente l'informazione, portare a cancellare il rapporto di lavoro.

Verifica della coerenza delle date

*** Assunzioni e cessazioni**

In caso di data inizio rapporto mancante o di data di cessazione antecedente all'inizio, il rapporto di lavoro è stato cancellato. Questa occorrenza, piuttosto marginale, ha interessato circa 4mila eventi.

*** Trasformazioni contrattuali**

Nel sistema sono presenti diverse trasformazioni contrattuali effettuate in date non coerenti con il periodo di vita del rapporto di lavoro. In questi casi lo schema di comportamento è il seguente:

- se la trasformazione avviene nell'intervallo compreso tra la data inizio del rapporto e il giorno successivo alla data di cessazione è valida ed in particolare:
 - se avviene nella data di cessazione o nel giorno successivo viene eliminata la cessazione
 - se avviene successivamente a questo termine (o antecedentemente alla data di inizio del rapporto) non viene considerata; il rapporto di lavoro quindi si conclude alla data originalmente.
- in tutti gli altri casi la trasformazione viene considerata nulla ed eliminata.

Verifica del tipo di trasformazione comunicato

Per valutare la correttezza delle trasformazioni contrattuali rispetto al rapporto di lavoro cui sono riferite è necessario conoscere:

- il contratto di assunzione
- il tipo di trasformazione comunicato e la sua compatibilità con il contratto di assunzione
- il contratto al momento dell'eventuale cessazione.

Sulla base di queste tre informazioni si determina la situazione in cui ricade la trasformazione esaminata e, di conseguenza, si definiscono le misure da intraprendere per correggere eventuali incoerenze. La gestione di queste situazioni avviene secondo lo schema di seguito riportato. Ne consegue che in alcuni casi di incompatibilità l'evento di trasformazione viene modificato in modo da risultare coerente con il rapporto di lavoro, in altri casi l'informazione portata dalla comunicazione di trasformazione permette di correggere gli altri eventi registrati (assunzione o cessazione) afferenti al medesimo rapporto di lavoro.

Contratto di assunzione	Trasformazione "concorde" con l'assunzione	Cessazione con contratto a tempo indet.	Soluzione
Cap, Ctd	Si	Si	Situazione corretta.
Cap, Ctd	Si	No	Correzione del contratto di cessazione (tempo indeterminato).
Cap, Ctd	No	Si	Correzione del tipo di trasformazione in base al contratto di assunzione.
Cap, Ctd	No	No	Correzione del contratto di cessazione (tempo indeterminato). Correzione del tipo di trasformazione in base al contratto di assunzione.
Som, Int, Dom, Par, Esp	No	-	Il contratto di cessazione viene sovrascritto dal contratto di assunzione. Cancellazione dell'evento di trasformazione.
Cti	No	-	Modificazione del contratto di assunzione in base al tipo di trasformazione.

Trasformazioni contrattuali implicite

In alcuni casi, il rapporto di lavoro può essere avviato con un determinato contratto (es. tempo determinato) e concludersi con un contratto diverso (es. tempo indeterminato), pur senza che sia stata comunicata una trasformazione contrattuale. In altre situazioni, alcuni rapporti di lavoro a termine (es. tempo determinato) non hanno una data di fine prevista valorizzata.

In generale, laddove l'evoluzione contrattuale risulta coerente con una possibile trasformazione "mancante", questa viene generata d'ufficio.

Verifica delle proroghe

Le comunicazioni di proroga devono attenersi ad alcune regole, per essere coerenti con il rapporto di lavoro a cui afferiscono, in particolare:

- le proroghe possono riguardare solo rapporti di lavoro a termine: tutte le proroghe applicate a contratti "non compatibili" (in generale rapporti a tempo indeterminato, anche se non strettamente dipendenti) verranno annullate.
- l'inizio della prima proroga deve avvenire contestualmente alla data comunicata di fine prevista. La fine dell'ultima proroga comunicata non deve superare il termine ultimo del rapporto (data di cessazione) o l'eventuale data di trasformazione contrattuale (verso il tempo indeterminato). Le comunicazioni di proroga che non rispettano questi vincoli verranno annullate.

Ricostruzione della data di cessazione

Nel caso in cui la mancanza dell'evento di cessazione non sia giustificabile, viene generata d'ufficio una chiusura del rapporto di lavoro e la durata è determinata ponendola pari alla moda delle durate per la medesima tipologia contrattuale. Questa situazione si verifica in un numero molto limitato di casi.

5. Data quality sulle storie lavorative

Il database statistico ha l'obiettivo di garantire, dove possibile, la coerenza dei rapporti di lavoro intesi come sequenza degli eventi che nel contesto informatico sono identificati dalle Co.

La messa in qualità del singolo rapporto di lavoro può essere estesa anche alla storia lavorativa di un singolo individuo, intesa come sequenza di rapporti di lavoro.

In questo modo, così come all'interno del singolo rapporto di lavoro non possono coesistere ad esempio più eventi di assunzione o di cessazione, nella carriera lavorativa di un soggetto non dovrebbero essere presenti rapporti di lavoro erroneamente sovrapposti o addirittura duplicati, se non in casi prestabiliti.

Rapporti di lavoro duplicati

Nonostante i controlli operati a monte dal Sil è possibile che nella storia lavorativa di un soggetto esistano due o più rapporti di lavoro apparentemente duplicati. La definizione di univocità di un rapporto di lavoro è legata ad alcuni aspetti chiave ossia: il lavoratore coinvolto, il datore di lavoro, la data di inizio del rapporto ed eventualmente la sede di lavoro ed il contratto. Consideriamo un rapporto come doppio, e quindi da eliminare, quando per la stessa coppia lavoratore/datore esiste un altro rapporto di lavoro avviato lo stesso giorno, con lo stesso contratto e nella stessa sede lavorativa.

Nell'ultima release del Puf sono stati eliminati circa 9mila rapporti di lavoro duplicati.

Overlapping

In presenza di sovrapposizione tra rapporti di lavoro per il medesimo lavoratore, prima di apportare delle modifiche è necessario accertarsi se l'overlapping sia legittimo o meno. Consideriamo la sovrapposizione "legittima" se:

- almeno uno dei due rapporti di lavoro è non subordinato/dipendente. Sono quindi ammissibili i casi in cui rapporti con contratto Intermittente, Parasubordinato, Domestico o Work experience si sovrappongono ad altri, oppure
- entrambi i rapporti di lavoro rientrano tra i contratti dipendenti (tempo indeterminato, apprendistato, tempo determinato, somministrato) e sono part-time.

Sono invece considerate non legittime, quindi da correggere, le situazioni in cui:

- entrambi i rapporti di lavoro sono subordinati (tempo indeterminato, apprendistato, tempo determinato, somministrato) e
- almeno uno dei due è a full-time.

Ad esempio:

- la sovrapposizione di un tempo indeterminato full-time con un rapporto parasubordinato è ammissibile,
- la sovrapposizione di due rapporti di lavoro a tempo determinato part-time è ammissibile,
- la sovrapposizione di un rapporto di lavoro a tempo indeterminato full-time con un altro a tempo determinato/apprendistato/somministrato non è ammissibile.

Nell'ultima release del Puf sono state individuate 453 **mila contratti che si sovrappongono illegittimamente a uno o più altri contratti**, interessando complessivamente circa **1 milione** di rapporti di lavoro (il 7% del totale) e **316 mila** lavoratori (il 10% del totale).

Le modalità di correzione sono state determinate a seconda del diverso tipo di sovrapposizione, individuato seguendo l'approccio di Allen per l'algebra degli intervalli.²

precedes	meets	overlaps	finished by	contains	starts	equals	started by	during	finishes	overlapped by	met by	preceded by
												
p	m	o	F	D	s	e	S	d	f	O	M	P

Table 1. Allen's thirteen basic relations

In linea generale, si procede per iterazioni successive individuando il rapporto di lavoro che causa il maggior numero di sovrapposizioni, all'interno della carriera di un lavoratore.

² Per maggiori informazioni, si veda <http://www.ics.uci.edu/~alspaugh/cls/shr/allen.html>

La data di inizio del rapporto non viene mai modificata, la procedura agisce solamente:

- “arretrando” la data di cessazione, nei casi meets, overlaps, finished by e starts (e reciproci) oppure
- cancellando il rapporto di lavoro, nei casi contains ed equals.

La procedura prosegue fino alla risoluzione di tutte le sovrapposizioni non consentite nella carriera lavorativa.

Ci sono poi una serie di eccezioni e particolarità dettate dalle caratteristiche del rapporto di lavoro (contratti, orari, ecc.).

Il risultato finale della procedura di correzione dell'overlapping riporta:

- **358mila** rapporti di lavoro con **data di fine modificata** (il 2,5% del totale)
- **95mila** rapporti di lavoro cancellati per **sovrapposizioni esatte o parziali** (lo 0,6% del totale)

6. Il contenuto di Mercurio

Tablelle dei fatti

- ft_datore_lavoro: dataset dei datori di lavoro
- ft_disoccupazione: eventi di disoccupazione (per calcolo di flussi e stock)
- ft_lavoratore: anagrafiche minime dei lavoratori
- ft_mobilità: eventi di mobilità dei lavoratori (entrate, sospensioni, uscite)
- ft_rapporti_lavoro: rapporti di lavoro

Tablelle di decodifica (o di appoggio)

- dt_ateco: classificazione dei settori di attività
- dt_cittadinanza: classificazione delle cittadinanze
- dt_contratto: classificazione dei contratti
- dt_dove: classificazione della geografia
- dt_evento_entrata_disoccupazione: classificazione degli eventi che determinano l'ingresso in disoccupazione
- dt_uscita_entrata_disoccupazione: classificazione degli eventi che determinano l'uscita dalla disoccupazione
- dtlivello_studio: classificazione dei livelli di istruzione
- dt_motivo_cessazione: classificazione dei motivi di cessazione
- dt_motivo_uscita_mobilità: classificazione dei motivi che determinano l'uscita dalla mobilità
- dt_nazione: classificazione delle nazioni
- dt_orario: classificazione degli orari di lavoro
- dt_qualifica: classificazione delle qualifiche/figure professionali
- dt_stato_entrata_disoccupazione: classificazione degli status occupazionali all'inizio del periodo di disoccupazione
- dt_trasformazione: classificazione delle tipologie di trasformazioni contrattuali
-

Alcune tablelle di appoggio, oltre alle codifiche “ministeriali”, propongono delle aggregazioni così come definite e utilizzate dall'Osservatorio di Veneto Lavoro.

Livello di aggiornamento delle anagrafiche

Le CO descrivono contenuti e situazioni di natura mutevole nel tempo: si pensi infatti alle caratteristiche dei lavoratori (quali il domicilio, la residenza di un lavoratore, il livello d'istruzione) oppure a quelle delle imprese (il settore di attività o l'ubicazione). E' probabile che nel periodo di vita del sistema informativo questi attributi siano variati almeno una volta. La nostra scelta è stata quella di riportare nel Puf il dato comunicato originariamente nella CO, ossia l'attributo che al momento della comunicazione era quello effettivo. Questa scelta consente di aggirare il problema della storicizzazione delle informazioni. Non siamo in grado di fornire una storicizzazione dei cambiamenti³ ma possiamo offrire qualità dell'informazione quando questa viene inviata in maniera puntuale; mantenendo i dati presenti nella CO originale riteniamo di offrire il massimo del dettaglio possibile.

Tracciato Record

Nel file Excel allegato è riportato il tracciato record delle tabelle dei fatti, con i riferimenti alle tabelle di appoggio.

7. L'acquisizione di Mercurio

La scelta dell'Osservatorio è stata di conservare per il Puf il formato utilizzato in tutte le elaborazioni. I dataset vengono quindi consegnati sottoforma di dump formato MySQL/MariaDB⁴ in un singolo file compresso.

Per creare le strutture e popolarle con i dati è necessario installare una versione server del DBMS MySQL/MariaDB; se non presente

- scaricare l'archivio del Puf e decomprimerlo per ottenere il file con estensione .sql
- creare un database vuoto nel proprio server MySQL/MariaDB
- eseguire il file .sql nell'ambiente di sviluppo MySQL/MariaDB5 per creare le tabelle e caricare i dati nel database creato.

Al fine di ridurre il volume di dati da trasferire, si è deciso di non indicizzare i campi chiave nelle tabelle fornite nel PUF. E' consigliabile creare degli indici nei campi di join utilizzati⁶.

³ Ricordiamo che non sono fonte di comunicazione obbligatoria i cambi di residenza e domicilio, il titolo di studio e molte altre variabili normalmente utilizzate nelle analisi del mercato del lavoro. Lasciamo quindi al ricercatore la possibilità di tracciare le ipotesi di cambiamento, mentre noi ci limitiamo a offrire le informazioni.

⁴ Come per le versioni precedenti del Puf, abbiamo ritenuto opportuno affidare la modalità di distribuzione ad un database open source. Si tratta in questo caso di MariaDB, controparte open source e non commerciale di MySQL. Per maggiori informazioni si veda <https://mariadb.org/> e <http://www.mysql.com/>.

⁵ Per maggiori informazioni sul caricamento di un dump sql, si veda <https://mariadb.com/kb/en/mariadb/backup-and-restore-overview/>.

⁶ Per maggiori informazioni sulla creazione e modifica di indici, si veda <https://mariadb.com/kb/en/mariadb/optimization-and-indexes/>

Riferimenti bibliografici

Prima dell'utilizzo del PUF, si consiglia vivamente di prendere visione dei seguenti documenti:

- Maurizio D., *Giove: un database statistico sul mercato del lavoro veneto*, collana “I tartufi”, n. 22, 2006. Link: <http://bit.ly/1qw7goW>
- Anastasia B., Disarò M., Gambuzza M. e Rasera M., *Comunicazioni obbligatorie e analisi congiunturale del mercato del lavoro: evoluzione, problemi metodologici, risultati*, collana “I tartufi”, n. 35, 2009. Link: <http://bit.ly/1VbxoBY>
- Anastasia B., Disarò M., Emireni G., Gambuzza M. e Rasera M., *Guida all'uso delle Comunicazioni Obbligatorie nel monitoraggio del mercato del lavoro. Seconda versione: dicembre 2010*, collana “I tartufi”, n. 36, 2010. Link: <http://bit.ly/1qvVlaZ>
- Anastasia B., Emireni G., Gambuzza M., Maschio S., Rasera M., *Grammatica delle comunicazioni obbligatorie /3 – Guida alle elaborazioni a partire dai dati di flusso*, collana “Tempi&metodi”, 2016. Link: <http://bit.ly/1NaSRmP>