Il processo di messa in qualità delle Comunicazioni Obbligatorie

Con la legge italiana n. 264 del 1949 si costituisce l'obbligo, per il datore di lavoro, di comunicare all'ufficio della PA di competenza l'avvio di un nuovo contratto di lavoro entro cinque giorni dall'inizio dello stesso. La legge, che inizialmente regolamentava solo il settore privato, è stata poi estesa anche al settore pubblico, prevedendo altre informazioni inerenti al rapporto stesso (es: tipologia contrattuale, salario etc.) e richiedendo la notifica obbligatoria anche per le variazioni e cessazioni del rapporto in oggetto. Tali comunicazioni attualmente sono note con il nome di "Comunicazioni Obbligatorie".

Negli anni la legge è stata ulteriormente estesa ed integrata fino a prevedere l'instaurazione di un archivio digitale volto all'osservazione delle dinamiche del mercato del lavoro, mediante la memorizzazione e l'analisi statistiche del database delle Comunicazioni Obbligatorie.

In questo documento si descrive il processo di messa in qualità dei dati delle Comunicazioni Obbligatorie, attraverso le quali siamo in grado di descrivere i principali eventi che caratterizzano l'evoluzione del mercato del lavoro: avviamenti al lavoro, cessazioni, proroghe di rapporti lavorativi esistenti o loro trasformazioni e dichiarazioni di immediata disponibilità (DID).

1. Le comunicazioni obbligatorie

A partire dall'anno 2008 (tramite la circolare No. 8371 del 21 Dicembre 2007 del Ministero del Lavoro) le Comunicazioni Obbligatorie, precedentemente inviate in formato cartaceo, vengono inviate in formato telematico ad un nodo di competenza per ciascuna regione. Una rete federata di nodi regionali e nazionali si occupa poi dell'instradamento delle comunicazioni ai nodi, la cui competenza è costituita sulla base di due principali fattori:

- la comunicazione riporta dati relativi ad un lavoratore domiciliato sul territorio di competenza del nodo;
- la comunicazione riporta dati relativi ad una sede operativa aziendale sul territorio di competenza del nodo.

La comunicazione obbligatoria riporta informazioni riferite al lavoratore, alla sede operativa della azienda presso cui viene instaurato il rapporto e al rapporto stesso. Per un maggiore dettaglio delle informazioni contenute all'interno della comunicazione si veda l'area riguardante le comunicazioni obbligatorie sul portale Cliclavoro (http://www.cliclavoro.gov.it/Aziende/Adempimenti/Pagine/Comunicazioni-Obbligatorie.aspx).

Le comunicazioni obbligatorie descrivono un dato di flusso che pertanto è finalizzato al monitoraggio degli eventi che avvengono nell'ambito del mercato del lavoro. In assenza di anagrafiche di riferimento o di dati di stock ciascuna comunicazione riporta interamente tutti i dati di interesse ed è autoconsistente. Le modalità stesse di invio e le tempistiche dello stesso non consentono validazioni dei contenuti delle comunicazioni se non a livello formale. Al momento della ricezione della comunicazione ciascun nodo può cioè valutare la consistenza interna della comunicazione (la corrispondenza delle classificazioni riportate ai vocabolari in uso, la consistenza delle date riportate all'interno della comunicazione), ma non valutare la consistenza con le altre comunicazioni ricevute (es: se una comunicazione altera un rapporto di lavoro non attivo). Il controllo

che viene effettuato è quindi di tipo sintattico, non semantico. Nel corso dell'adozione delle comunicazioni telematiche inoltre sono progressivamente stati inseriti nuovi controlli all'interno del processo che hanno portato ad un progressivo miglioramento della qualità del dato raccolto.

Il formato delle Comunicazioni Obbligatorie varia nel tempo così come il loro contenuto in funzione di diversi possibili eventi:

- Il cambiamento delle classificazioni adottate;
- cambiamenti normativi che comportano modifiche sulle modalità di raccolta dei dati e sui loro contenuti.

Di conseguenza nel tempo cambiano le regole da applicare per la verifica dei contenuti delle comunicazioni e come si vedrà nel seguito anche i processi di messa in qualità dell'informazione.

2. Dal dato amministrativo al dato statistico

Le Comunicazioni Obbligatorie rappresentano a tutti gli effetti un dato amministrativo: l'informazione viene raccolta per adempiere a requisiti normativi e viene utilizzata per la verifica degli eventi a cui fa riferimento. Come altri dati di tipo amministrativo l'informazione raccolta non può quindi essere modificata in quanto rappresenta a tutti gli effetti una comunicazione ufficiale; inoltre il valore della comunicazione è puntuale e l'interesse è proprio rivolto alla descrizione del singolo evento.

Il processo di messa in qualità che ci apprestiamo a descrivere è finalizzato alla definizione di un dato statistico: un dato cioè finalizzato all'indagine di un fenomeno che prescinde dunque dai singoli eventi che lo compongono. L'interesse è rivolto non tanto alla descrizione puntuale del singolo evento quanto alla descrizione dei fenomeni e degli andamenti che l'insieme di tali eventi determina. Proprio per questo l'attenzione è rivolta non tanto alla correttezza puntuale della singola comunicazione quanto alla coerenza dell'insieme delle comunicazioni e alla validità delle relazioni tra di esse. Per poter raggiungere questo obiettivo è lecito apportare modifiche all'informazione al fine di aumentarne la qualità complessiva e la sua capacità descrittiva dei fenomeni.

Un semplice esempio può chiarire la differenza tra i due tipi di dato e tra i trattamenti a cui vengono sottoposti: nel caso in cui venga comunicata la cessazione riferita ad un rapporto di lavoro non attivo, dal punto di vista amministrativo non emergono problemi, purché il contenuto della comunicazione sia corretto e coerente. Dal punto di vista statistico invece l'informazione non è consistente, mancando il corrispondente avviamento al lavoro. Dal punto di vista amministrativo non è dunque lecito generare una comunicazione di avviamento, in quanto creerebbe una comunicazione di fatto non avvenuta; dal punto di vista statistico è invece necessario creare il corrispondente avviamento per garantire la coerenza del rapporto in esame.

In conclusione il dato amministrativo può essere interpretato come puntuale ed autoconsistente, su di esso si possono condurre analisi di flusso aggregate, ma esse riguarderanno comunque una distribuzione di eventi puntuali.

Il dato statistico invece deve essere sempre considerato nel suo complesso come parte di un insieme di eventi che devono mantenere una consistenza anche a livello globale (la successione di eventi in una carriera

ad esempio deve essere sensata, non si possono avere solo avviamenti al lavoro senza nessuna cessazione) e come tale deve essere trattato: l'arrivo di una nuova comunicazione comporta dunque l'aggiornamento di un insieme di informazioni storicizzate considerate nel loro complesso.

3. Criticità del processo di messa in qualità

Il processo di messa in qualità presenta alcune criticità che è doveroso puntualizzare poiché hanno inciso sulle soluzioni adottate:

- Il dato amministrativo alla base del processo deve garantire la correttezza dell'informazione: le informazioni riportate all'interno di ciascuna comunicazione devono essere quanto più complete possibile e devono rispondere ai requisiti formali delle Comunicazioni Obbligatorie in termini di:
  - formati (ad esempio le date devono essere nei formati corretti);
  - contenuti (ciascun campo deve contenere l'informazione corretta, ad esempio un campo data deve contenere una data e non un numero);
  - completezza (i campi devono essere quanto più possibile valorizzati);
  - vocabolario (il contenuto di un campo riconducibile ad una classificazione deve riportare un valore appartenente alla classificazione stessa).
- Il formato dei dati può variare nel tempo, sono quindi necessari meccanismi che consentano di validare la comunicazione in funzione del momento in cui viene ricevuta, così come meccanismi che permettano di ricondurre i diversi formati ad un unico formato finale di analisi (solitamente quello corrente).
- Il dato amministrativo deve essere trasformato in dato statistico: oltre alla correttezza interna deve essere garantita anche la coerenza tra le diverse comunicazioni. Devono cioè essere previsti processi che analizzino la sequenza delle comunicazioni, ne verifichino la coerenza e intervengano dove necessario per correggere eventuali errori.
- La comunicazione oltre che sul presente (con la registrazione della comunicazione) e sul futuro (con la registrazione di date di cessazione previste per i rapporti a tempo determinato) può avere effetti anche sul passato: il processo di messa in qualità deve poter intervenire anche sul dato storico, se necessario, per garantire la sequenza delle informazioni (es: inserendo un avviamento passato non presente al momento dell'arrivo della relativa cessazione).

L'obiettivo dell'intero processo è quello di disporre di un insieme di informazioni coerenti che consentano non solo di ricostruire il flusso in modo corretto, ma anche di ricostruire le carriere lavorative (almeno per la porzione oggetto di comunicazione) e le anagrafiche dei soggetti interessati (lavoratori e aziende) contribuendo alla progressiva costruzione di uno stock.

4. Modello dati di riferimento

In funzione delle considerazioni espresse in precedenza e per poter comprendere il modello dati utilizzato all'interno del processo, è necessario introdurre alcuni concetti ed assiomi su cui si è basato il modello dati utilizzato.

- **Evento**: una Comunicazione Obbligatoria è modellata come un evento osservato in un momento temporale definito. Per poter usufruire di un modello flessibile un evento può essere di qualsiasi tipo: un avviamento al lavoro, una trasformazione, una proroga, una cessazione, una DID. L'evento è l'elemento base su cui si fonda l'intero modello e la maggior parte delle informazioni provenienti dal sistema alimentante vengono ricondotte a tale concetto. Un evento è caratterizzato da una data di inizio e da uno o più soggetti interessati (persone, imprese, ecc.).
- **Rapporto**: gli eventi possono essere aggregati in rapporti: tutti gli eventi successivi e contigui che legano due soggetti (lavoratore ed azienda, ad esempio la filiera avviamento, proroga, trasformazione, cessazione) concorrono alla creazione di un unico rapporto di lavoro. Il rapporto rappresenta il massimo livello di aggregazione degli eventi e il punto di partenza per tutte le aggregazioni successive.
- **Transizione**: due rapporti legati da successione temporale concorrono a definire una transizione, cioè un passaggio da un rapporto ad un altro. Le transizioni hanno particolare importanza nello studio delle evoluzioni dei rapporti e di conseguenza delle carriere.
- **Persone**: le persone rappresentano una delle tipologie di soggetti che possono essere interessate da eventi. Le persone possono essere dettagliate in lavoratori, studenti, ecc. ma tutte le diverse accezioni mantengono una serie di caratteristiche comuni come i dati anagrafici, la carriera ecc. Per ciascuna persona viene conservato lo storico dei dati passibili di variazione nel tempo.
- **Imprese**: una seconda tipologia di soggetti interessata da eventi è quella delle imprese. Esse possono, per mezzo degli eventi, rapportarsi alle persone e stabilire con loro dei rapporti. Per ciascuna impresa viene conservato lo storico dei dati suscettibili di variazione nel tempo.

Come detto in precedenza l'evento rappresenta l'elemento base su cui si fonda l'intero modello. Gli eventi vengono caricati a partire dalle fonti informative disponibili riconducendole ad un modello dati comune in grado di registrare le caratteristiche salienti di ciascun evento minimizzando la perdita in termini informativi e consentendo nel frattempo di confrontare e combinare fra loro eventi in prima analisi differenti.

Attraverso il processo di aggregazione gli eventi vengono tradotti in rapporti cioè in associazioni tra due soggetti (tipicamente una persona e un'impresa o un ente) aventi caratteristiche distintive ed un periodo di validità. I rapporti riferiti ad un medesimo soggetto devono essere omogenei fra di loro evitando sovrapposizioni temporali se non nei casi espressamente previsti (ad es. part time).

Le transizioni vengono costruite associando rapporti contigui riferiti ad un medesimo soggetto evidenziando di fatto il passaggio da uno stato al successivo nella successione temporale; gli stock (annuali inizialmente ma non sono escluse successive aggregazioni in periodi di interesse differenti) vengono generati filtrando i rapporti in base al periodo temporale di interesse ed inserendo nel medesimo stock quelli riguardanti la finestra temporale di interesse. In caso di rapporti estesi su più periodi essi vengono segmentati in più sotto rapporti ciascuno dei quali viene associato al corretto intervallo temporale.

Ancora una volta a partire dai rapporti, ma operando in questo caso aggregazioni di tipo logico, vengono costruite le strutture riguardanti persone e imprese.

Aggregando gli eventi in base ai soggetti principali di interesse è possibile ricavare l'elenco delle persone interessate. Una persona entra a far parte di questo insieme se esiste almeno un evento che la riguarda tra quelli registrati in banca dati. Come è logico, l'archivio che ne consegue ha carattere incrementale e non riguarda, almeno per un periodo transitorio iniziale, l'intero universo delle persone. A partire dagli eventi è

possibile generare alcune strutture di corredo dell'entità persona: i dati storici registrano i cambiamenti nei dati anagrafici e di domicilio avvenuti nel corso del tempo per permettere di condurre analisi retrospettive considerando non solo la condizione attuale del soggetto ma la sua situazione al momento dell'evento considerato; lo stock annuale riassume attraverso una serie di indicatori la situazione del soggetto nel corso dell'intero periodo di interesse prescindendo dai singoli rapporti. Esistono infine alcune informazioni aggiuntive che, una volta ricavato il soggetto, possono essere associate ad esso a partire da fonti informative esterne. Nell'utilizzare tali informazioni è necessario porre particolare attenzione alla granularità temporale dell'informazione: non sempre infatti è possibile associare qualsiasi informazione poiché alcune di esse hanno carattere puntuale, altre ad esempio annuale; è quindi fondamentale prima di mettere in relazione dati riguardanti un medesimo soggetto provenienti da fonti diverse verificare che l'aggregazione temporale sia la medesima ed eventualmente procedere all'aggregazione del dato di dettaglio.

Analogamente a quanto avviene per la persona, le imprese possono essere ricavate aggregando gli eventi in base al secondo soggetto di interesse. Il meccanismo, al netto delle differenze in termini di contenuti, è analogo a quello adottato nella ricostruzione delle persone e porta alla definizione di strutture accessorie come l'anagrafica, gli stock annuali e i dati storici. L'integrazione di ulteriori fonti informative può portare all'associazioni di informazioni di particolare interesse riguardanti l'impresa nel corso degli anni.

5. Anonimizzazione dei dati

Il passaggio dal dato amministrativo al dato statistico prevede lo spostamento dell'attenzione dal singolo soggetto ai fenomeni che lo riguardano. In tal senso non è più di interesse poter riconoscere il singolo soggetto, ma solo poterlo identificare univocamente all'interno degli archivi. Per poter quindi mantenere tale tracciabilità astraendo dall'identificazione puntuale, è stato adottato un algoritmo di anonimizzazione delle informazioni identificative (codice fiscale dei soggetti e partite IVA delle aziende) basato su una codifica ottenuta tramite un algoritmo di hashing. Tutte le comunicazioni caricate vengono sottoposte alla medesima procedura in modo da assicurare la possibilità di ricondurre le informazioni riferite al medesimo soggetto.

L'algoritmo di hashing utilizzato è di tipo SHA-1. Per maggiori dettagli riguardo le sue specifiche si veda: http://www.faqs.org/rfcs/rfc3174.html. L'algoritmo è stato implementato a partire da una versione già testata, disponibile all'indirizzo: http://pajhome.org.uk/index.html opportunamente adattata per poter essere utilizzata in modo indipendente ed ottimizzata per carichi elevati di lavoro.

L'utilizzo di tale algoritmo garantisce quindi l'anonimato dei soggetti analizzati preservandone la tracciabilità all'interno delle banche dati.

6. Conclusioni

In questo documento di sintesi si è descritta una metodologia unificata, ripetibile ed aperta per l'analisi e la messa in qualità dei dati, mostrandone l'applicazione nel dominio delle Comunicazioni Obbligatorie del Mercato del Lavoro.

L'applicazione dell'ETL sul database del Mercato del Lavoro ha permesso di analizzare la qualità del dato sorgente e generare un nuovo dataset statistico qualitativamente superiore nei termini di accuratezza, consistenza e completezza.

Le tecniche utilizzate all'interno del processo, hanno evidenziato la forte dipendenza che connette la qualità del dato in ingresso con la sua valorizzazione statistica: all'aumentare della qualità del primo aumenta l'efficacia che le informazioni statistiche da esso derivate hanno nel processo decisionale.