Data Quality: Un Approccio Metodologico ed Applicativo

"Il caso delle COB del mercato del lavoro in Lombardia"

[WORKING PAPER]

Documentazione relativa alla ricostruzione di una metodologia unificata, ripetibile e aperta

A cura di Mirko Cesarini, Matteo Fontana, Fabio Mercorio,
Mario Mezzanzanica, Nicolò Vegetti

Il presente lavoro è frutto della collaborazione di un progetto di ricerca tra ARIFL *(Agenzia Regionale per l'Istruzione, la Formazione e il Lavoro)* ed il CRISP *(Centro di Ricerca Interuniversitario per i Servizi di Pubblica utilità*) dell'Università di Milano-Bicocca.

Indice

1. Introduzione al Problema del Data Quality

La società attuale è caratterizzata da una forte pervasività dell'informazione digitale. Basti pensare alle applicazioni che supportano le attività di gestione dei processi operativi aziendali, come pure alle applicazioni che gestiscono i processi della Pubblica Amministrazione (PA), sempre più presenti sia internamente alle organizzazioni sia nel rapporto con i clienti, cittadini ed imprese.

Nell'ultimo decennio, quindi, è cresciuta l'esigenza di analizzare la grande mole di dati presenti negli archivi digitali. Il contenuto informativo che questi archivi forniscono, inteso come l'informazione utile al processo decisionale che può essere derivata da essi, è spesso limitato dalla difficoltà d'integrare diverse fonti (a volte eterogenee tra loro) e dal basso livello qualitativo dei dati stessi. Per tali ragioni, lo sviluppo di metodi e tecniche di integrazione ed analisi della qualità dei dati è divenuto un fattore cruciale sia per il supporto dei processi decisionali, sia per servizi web volti all'interazione tra aziende/PA e il cittadino (noti rispettivamente come *eBusiness* ed *eGovernment*).

Il termine *Data Quality* è genericamente usato per descrivere un processo di analisi cui sottoporre i dati, con il fine di analizzarne ed incrementarne la qualità. In funzione della natura del dato e dello scopo per il quale esso viene analizzato, il termine "qualità" viene spesso declinato nei concetti di *accuratezza*, *consistenza*, *completezza*, *correttezza* (ossia, le *dimensioni* della qualità) che ne definiscono le proprietà generali. La comunità scientifica ha definito una grande varietà di dimensioni, per un approfondimento in proposito si rimanda il lettore al lavoro di (Batini & Scannapieco, 2006).

L'analisi, ed il successivo processo volto alla messa in qualità dei dati, sono divenute nel tempo attività propedeutiche al supporto decisionale. Infatti, il contenuto informativo che il dato può esprimere è funzionale alla sua capacità di descrivere l'ambiente dal quale è stato tratto o osservato. E' ben noto che l'utilizzo di dati di bassa qualità può causare decisioni errate o inefficienti, determinando danni all'economia dell'azienda o dell'amministrazione pubblica così come a tutti i soggetti che su tali analisi basano i loro processi decisionali. Il problema del data quality gioca un ruolo determinante in moltissimi contesti: scientifico, sociale, economico, politico etc. A tal proposito, si ricorda un caso eclatante verificatosi nel campo dell'esplorazione spaziale: l'esplosione dello Space Shuttle Challenger è stata imputata a dieci differenti categorie di problemi di data quality (come apliamente descritto nel lavoro di (Fisher & Kingma, 2001)). Un ben più noto caso è quello del *Millenium Bug*, che ha ottenuto un grande risalto mediatico soprattutto a causa dell'elevato impatto economico che ne è derivato, sia nel settore pubblico che nel privato. Nonostante ci siano stime discordanti sui costi che questo problema di data quality ha avuto sulle diverse economie industrializzate, si concorda nel riconoscere un costo per la messa in qualità dei sistemi prima dell'anno 2000 almeno pari a 400 miliardi di dollari statunitensi. In letteratura sono diversi gli studi che analizzano le conseguenze della scarsa qualità dei dati, di cui soffrono molte industrie, siano esse private o pubbliche amministrazioni, tra i lavori più celebri si ricordano (Redman T. C., 1998) (Strong, Lee, & Wang, 1997). Le relative comunità scientifiche hanno sviluppato diverse tecniche per

l'analisi e il miglioramento della qualità dei dati, ben declinando le varie dimensioni che la definiscono.

Attualmente, una sfida nell'ambito del data quality risiede nello sviluppo di metodologie e strumenti capaci di analizzare e migliorare la qualità dei dati, così da massimizzarne la capacità informativa. Chiaramente, è cruciale lo sviluppo di metodologie scalabili, cioè in grado di gestire efficientemente una grande mole di dati (es., *Big Data*), che sovente è presente in molti contesti aziendali e della PA.

Per meglio chiarire il problema, si osservi la tabella 1 che riporta, a titolo esemplificativo, un registro di un armatore nel quale vengono memorizzati gli ingressi (checkin) e le uscite (checkout) che una nave effettua ogni volta che attracca/salpa da un porto di transito. È evidente che la data di partenza dal porto di Lisbona è mancante dal momento che la nave non può entrare in un porto senza aver prima salpato da quello di partenza. Questa situazione identifica una *inconsistenza* nel database.

Tabella 1

ShipID	*Città*	*Data*	*Tipo di Evento*
S01	Venezia	12 Aprile 2011	checkin
S01	Venezia	15 Aprile2011	checkout
S01	Lisbona	30 Aprile2011	checkin
S01	Barcellona	5 Maggio 2011	checkin
S01	Barcellona	8 Maggio 2011	checkout
...			

Si immagini ora di voler calcolare l'indicatore "*giorni di navigazione"* su un dataset come quello mostrato in Tabella. Il dato mancate (la partenza da Lisbona) può essere inputato osservando che le navi rimangono in porto per circa 3 giorni, sulla base di altre informazioni note all'armatore, oppure osservando i termini della legislazione delle navi mercantili, etc. Si noti che un'inconsistenza come quella appena descritta può avere un impatto sensibile sui dati aggregati, così come può essere trascurabile in proporzione alla frequenza con la quale questa appare ripetutamente nel dato sorgente. In ogni caso, l'individuazione di problemi di qualità (e la formalizzazione del processo con il quale queste vengono corrette) è cruciale affinchè l'informazione derivata/aggregata sia quanto più veritiera possibile e, quindi, realmente utile al processo decisionale.

2. Data Quality

2.1 Alcune dimensioni del Data Quality

Il termine Data Quality identifica genericamente attività e processi volti all'analisi (ed eventuale miglioramento) della qualità dei dati di un database. Tuttavia, la qualità di un dato può essere osservata ponendo l'accento su alcuni aspetti che, per l'esperto di dominio, possono risultare più rilevanti di altri. A tal fine, le *dimensioni* del data quality si propongono come strumento (qualitativo) per la valutazione della qualità dei dati. E' importante osservare che tali dimensioni, alcuni delle quali saranno introdotte di seguito, possono essere definite a livello di *schema*, di *processo* o di *dato*.

Nel primo caso si analizza la struttura logica utilizzata per rappresentare il dato, con lo scopo di verificare che sia adeguata ed idonea ad ottenere un dato con le caratteristiche di qualità richieste. Si immagini, ad esempio, di voler rappresentare i dati anagrafici dei cittadini di un comune. Una struttura logica che permetta l'inserimento non diversificato di nome e cognome (il valore di nome e cognome in un campo unico piuttosto che in due campi distinti) non risulterà adeguata per una rappresentazione qualitativa del dato in oggetto, poiché si presterà maggiormente ad errori, omonimie ed incogruenze. Similmente, l'analisi a livello di processo verifica che il procedimento con cui il dato viene osservato o raccolto sia idoneo, nella fattispecie si analizza la modalità con la quale i dati anagrafici vengono raccolti. Nell'esempio appena esposto, l'impiegato comunale potrebbe inserire solo il nome, senza il cognome, o viceversa, generando un evidente problema di qualità. Nell'analisi a *livello di dato*, invece, si analizza direttamente il dato memorizzato, astraendosi dalla forma e dal modo con il quale è pervenuto. Nel nostro esempio, un nome contenente cifre numeriche è indice di un problema di data quality nel record trattato.

È importante sottolineare che la qualità a livello di schema incide sulla qualità a livello di processo, che a sua volta ha impatto sulla qualità del dato finale. Tuttavia, non sempre è possibile analizzare/correggere anomalie qualitative a livello di schema o di processo (es., lo schema non è accessibile, il processo non è osservabile né alterabile). In questi casi, l'analisi a livello di dato quindi rimane l'unica alternativa percorribile.

Di seguito, e per il resto di questo documento, si farà riferimento a dimensioni del data quality a livello di dato e ci si limiterà alla descrizione delle dimensioni della qualità di interesse per il caso in oggetto, per approfondimenti si rimanda il lettore al lavoro di (Redman T. C., 2001).

2.2 Accuratezza

L'Accuratezza del dato è definita come la distanza tra un valore **v** e un valore **v'** considerato come la corretta rappresentazione del fenomeno reale che v intende esprimere.

Si consideri, a titolo esemplificativo, un valore di un attributo testuale **v** che contiene un nome proprio di persona, e che l'insieme di tutti i possibili nomi ammessi, in questo specifico esempio (ossia, il dominio D) sia noto e limitato (un dominio finito). Ad esempio,

si potrà avere l'attributo valorizzato con **v** = "Andra". Si potrà valutare l'accuratezza di **v** ponendo l'attenzione su:

- **Accurratezza Sintattica**. Si verifica che il valore dell'attributo **v** sia presente nell'insieme dei valori di dominio D. In altri termini, nell'esempio in oggetto, si verifica che "Andra" sia un nome proprio contenuto nel nostro dizionario D. È facile immaginare che tale valore non risulterà presente in D, e si potranno quindi ottenere dei valori *vicini* (in questo caso la vicinanza può essere realizzata come il numero di lettere necessarie per rendere i due valori uguali, ma si possono definire altre metriche). Si potrà scoprire, ad esempio, che un valore ammissibile è "Andrea", come pure "Alessandra", e che il primo è più vicino a **v** di quanto non lo sia il secondo. Nel caso dell'accuratezza sintattica, però, non si è interessati alla valutazione del valore **v** con il valore reale **v'** (cioè con il vero nome dell'individuo che si vuole rappresentare) ma con l'insieme di tutti i valori di dominio dell'attributo **v**. In tal caso, poiché "Andrea" è il valore più vicino a **v**, sarà quest'ultimo il valore che sarà usato per determinare l'accuratezza del valore **v** = "Andra".
- **Accuratezza Semantica**. In questo caso si valuta l'accuratezza del valore **v** paragonandolo con la sua controparte reale **v'**. È chiaro che è fondamentale conoscere qual è il vero nome dell'individuo che l'attributo v vuole esprimere. Si potrebbe scroprire, ad esempio, che il valore reale **v'** è "Alessandro", piuttosto che "Andrea". Diversamente dall'accuratezza sintattica, che misura la distanza tra valore osservato e valore reale come valore numerico, l'accuratezza semantica fornisce una valutazione dicotomica: o **v** è accurato quanto il valore reale o non lo è, indipendentemente dalla distanza tra i valori **v** e **v'**. Come conseguenza, grazie all'accuratezza semantica, si esprime intrinsecamente il concetto di **correttezza** del dato.

2.3 Completezza

In letteratura, la completezza è definita come "il livello di ampiezza, profondità ed appropriatezza di un dato in funzione dello scopo che ha." (Wang & Strong, 1996).

Per meglio descrivere la dimensione della completezza, a titolo esemplificativo, si può immaginare la struttura che memorizza i dati come una tabella (relazione): le *colonne* rappresentano gli attributi dell'oggetto che si vuole memorizzare, mentre le *righe* della tabella (tuple) rappresentano le diverse osservazioni dell'oggetto. Ad esempio, nel caso dell'anagrafica comunale, si può immaginare il dato sulla popolazione anagrafica come una tabella in cui le colonne modellano le informazioni anagrafiche dei cittadini (es., nome, cognome, sesso, data di nascita, etc) mentre ogni riga rappresenta un cittadino diverso.

È possibile quindi distinguere tra diverse tipologie di completezza del dato:

- La completezza di *colonna*, che misura la mancanza di specifici attributi o colonne da una tabella;
- la completezza di *popolazione*, che invece analizza le tuple mancanti in riferimento ad una popolazione osservata.

Risulta evidente che alcuni livelli di completezza sono difficili da valutare. Ad esempio, nel caso di una relazione contenente dati anagrafici di un cittadino, la mancanza di uno o

più valori per l'attributo *data di nascita* rappresenta una chiara incompletezza del dato in termini di colonna.

Diversamente, se volessimo calcolare quanto il nostro dataset è completo in termini di rappresentatività della popolazione, sarà necessario conoscere con esattezza il valore della popolazione di riferimento. A tal proposito, si immagini di voler misurare la completezza di popolazione di un database contenente i dati di dei giovani *neet* (not in education, employment or training) italiani. Nonostante le caratteristiche della classe siano chiare, non è facile ottenere la valutazione della completezza di un tale database poichè è difficile individuare il valore esatto della popolazione attuale dei *neet.*

Generalmente, la completezza viene espressa in termini di tasso di completezza, calcolato come il rapporto tra la cardinalità del campione in possesso e la cardinalità della popolazione.

2.4 Consistenza

La consistenza è definita in letteratura, in riferimento alla *"violazione di una o più regole semantiche definite su un insieme di dati"* (Batini & Scannapieco, 2006). Anche in questo caso, è possibile identificare vari livelli di consistenza:

- **Consistenza di chiave**: è la più semplice delle forme di consistenza e richiede che, all'interno di uno schema di relazione (una tabella), non vi siano due tuple con il medesimo valore di un attributo usato come chiave. Nell'esempio dell'anagrafica si potrebbe usare il campo *codicefiscale* come chiave. In tal caso la cosistenza di chiave richiederebbe che non ci siano due persone con lo stesso codice fiscale. Una perfetta omonimia di nomi, date e luoghi di nascita, seppur estremamente improbabile, violerebbe questo vincolo di consistenza, mostrando l'inadeguatezza del campo *codicefiscale* a svolgere tale compito.
- **Consistenza di inclusione**: ne è un classico esempio è la "foreign key" di una relazione. Richiede che un insieme di colonne di uno schema di relazione sia contenuto in un altro insieme di colonne dello stesso schema di relazione, o di un'altra istanza di schema di relazione. In riferimento al nostro esempio, si immagini di avere, oltre alla tabella *anagrafica*, anche una seconda tabella in cui sono memorizzati i dati sui nuclei familiari, identificati univocamente dal codice fiscale del capofamiglia. È chiaro che esiste una relazione tra i dati delle due tabelle. Potremmo dire, infatti, che le tabelle *nucleifamiliari* ed *anagrafica* sono in relazione tra loro attraverso il campo *codicefiscale* del cittadino capofamiglia, che rappresenza la "foreign key" della relazione. Affinchè la consistenza di inclusione (noto anche come *vincolo di integrità referenziale*) sia soddisfatta, tutti i capofamiglia di un nucleo familiare dovranno essere presenti nella tabella *anagrafica.* Non può esistere, infatti, un capofamiglia che non sia anche un cittadino regolarmente iscritto nella tabella *anagrafica*. Contrariamente, può esistere un cittadino il cui codice fiscale non sia elencato nella tabella dei *capofamiglia.*
- **Dipendenze funzionali**: sono le più note e le più utilizzate. In generale, data una relazione R con gli attributi X e Y, si dice che Y è funzionalmente dipendente da X (X->Y) se e solo se per ogni valore di X è associato un preciso valore in Y. In altri

termini, data una tupla con un valore di X la dipendenza funzionale esprime la capacità di conoscere con certezza il valore dell'attributo Y. Nell'esempio dell'anagrafica, il campo codicefiscale è funzionalmente dipendente dai campi necessari per il suo calcolo (ossia, *nome, cognome, data di nascita, luogo di nascita, sesso*). Infatti, una volta noti questi campi, è possibile generare uno ed un solo codice fiscale ad essi associato.

La definizione di consistenza è generica e permette quindi la modellazione di una grande quantità di "regole semantiche". E' possibile, infine, identificare regole semantiche che gli approcci appena descritti non possono esprimere (Il registro dell'armatore navale precedentemente descritto ne è un esempio) per i quali è necessario realizzare delle soluzioni ad-hoc, come sarà descritto in seguito.

2.5 Una Panoramica sullo Stato dell'Arte

Da un punto di vista di ricerca scientifica, il problema della qualità dei dati è stato affrontato in diversi contensti, inclusi quello statistico, economico e non ultimo quello informatico. Di seguito si descriveranno i principali approcci al problema dell'analisi e messa in qualità dei dati attinenti agli argomenti trattati in questo documento, organizzandoli per *macro* approcci:

- **Verifica e Miglioramento del Processo di Raccolta Dati**: Consiste nell'analisi approfondita dell'intero processo di raccolta del dato col fine di individuare e migliorare le fasi del processo che minano la buona qualità del dato (es., modificare il processo di inserimento dati da manuale ad automatico). A tal proposito è importante sottolineare due aspetti: (1) In molti contesti reali non è possibile né intervenire né analizzare il processo di raccolta. (2) L'invertento, con eventuale miglioramento, del processo di raccolta non fornisce alcuna garanzia sulla qualità del dato: è sempre necessaria un'analisi della qualità che certifichi come l'intervento sul processo di raccolta abbia prodotto, come effetto, un incremento della qualità del dato raccolto.
- **Integrazione:** Prevede l'integrazione del dato attraverso un confronto con le *controparti reali*. Tuttavia, poiché questo metodo è estremamente oneroso, sia in termini economici che in termini temporali, la sua applicazione è limitata a database di dimensioni ridotte.
- **Record Linkage:** *(noto anche come object identification, record matching, merge-purge problem)* consiste nel confronto tra record che contengono informazioni sullo stesso oggetto ma che provengono da database differenti (es., si incrociano i dati di sorgenti diverse). Si assuma di avere i dati anagrafici e commerciali di un insieme di aziende in due database distinti ed isolati (es., i database sono stati manutenuti da filiali diverse della stessa impresa). È possibile combinare i due database in uno unico derivando un maggior numero di informazioni per ogni singola azienda. I dati combacianti si presuppongono corretti, mentre quelli differenti vengono segnalati per successive verifiche. La tecnica descritta èmolto efficace, ma applicabile sono nei casi in cui è possibile ottenere due dataset di fonti diverse che descrivono la stessa realtà. Inoltre, è necessario avere

una *chiave (link)* che permetta di associare uno o più record dei differenti database. Sfortunatamente non sempre è possibile avere dei database con queste caratteristiche. In questi casi, la tecnica può essere applicata ma è necessario svolgere (1) il prodotto cartesiano dei domini dei due dataset sorgenti (ossia, si generano tutti i possibili modi di associare tutti gli elementi di un database con tutti gli elementi dell'altro), ottendendo uno spazio di grandi dimensioni che dovrà essere ulteriormente raffinato; (2) successivamente si procede individuando un sottoinsieme dello spazio di ricerca sul quale focalizzare l'attenzione, (3) infine si definisce un modello decisionale (es., una funzione "distanza") che definisce se due record dei due database distinti sono correlati o meno. Come è facile immaginare, l'applicabilità dell'approccio è limitata dalle dimensioni dello spazio di ricerca e dalla corretta calibrazione del modello decisionale, sulla base del quale si possono ottenere dei falsi positivi/negati. Storicamente, la tecnica del record linkage è stata dapprima descritta da (Dunn, 1946), successivamente è stata estesa ed ampliata attraverso lo sviluppo di diverse approcci (es., probabilistici, euristici, logic-based etc), per approfondimenti in merito si suggerisce (Fan, 2008).

- **Basati su regole e vincoli di dominio:** In molte applicazioni reali l'approccio basato sull'integrazione dei dati così come sul record matching è inapplicabile, oppure troppo oneroso in termini computazionali. In questi casi si usa un metodo basato su ispezione e correzione. È l'approccio più utilizzato e, quindi, quello che presenta il maggior numero di varianti e sviluppi. In alcuni casi sono state sviluppate delle tecniche formali o algoritmi, in altri dei veri e propri tool (commerciali e non). Di seguito, per chiarezza espositiva, si distinguerà tra approcci basati (1) su *regole derivate da conoscenza del dominio* (es: gli esperti del dominio definiscono algoritmi e tool ad-hoc per l'analisi e la messa in qualità dei dati), (2) sulle *dipendenze* (formalizzate da esperti di dominio) e quelli (3) basati sull'*apprendimento* (ossia, un software capace di apprendere una proprietà sulla base di un dataset di riferimento)*:*
  - **Basati sulle dipendenze:** Una dipendenza funzionale tra due attributi di uno schema di relazione permette di associare con certezza il valore di un attributo al valore di un altro. Tuttavia, le dipendenze funzionali possono esprimere solo vincoli su l'intero set di attributi. Chiaramente, nel corso degli anni le diverse comunità scientifiche hanno sviluppato molte varianti delle dipedenze funzionali, tra cui le multivalued dependencies (Fagin, 1977) e le embedded multivalued dependencies (Beeri, Fagin, & Howard, 1977), fino alle recentissime Conditional Functional Dependencies (Fan, 2008), la cui trattazione esula dal contesto di questo documento. Tuttavia, è ben noto che esistono vincoli semantici che le dipendenze funzionali in genere non possono esprimere, come già provato da (Vardi, 1987).

    Un diverso uso delle dipendenze è nell'approccio volto a individuare un "repair" del database, cioè un nuovo database che sia consistente rispetto all'originale (Chomicki & Marcinkowski, 2005). E' un approccio assolutamente promettente, la cui scalabilità tuttavia non è ancora stata analizzata su database di grandi dimensioni con un gran numero di vincoli di qualità. Tuttavia, in alcuni casi i problemi di data quality possono verificarsi frequentemente, soprattutto nei database incrementali. Per gestire queste delicate situazioni la comunità scientifica ha sviluppato la tecnica del

Consistent Query Answering (Arenas, Bertossi, & Chomicki, 1999) capace di gestire problemi di consistenza dei dati on-demand: si calcola una risposta consistente ad una query effettuata su un dato inconsistente *lasciando inalterato il dato originale*. In questi termini, una risposta è considerata consistente quando i dati, a partire dai quali la risposta è elaborata, appaiono in tutte le possibili versioni (repair) del database. Anche in questo caso, l'applicabilità dell'approccio è limitata dall'espressività delle dipendenze funzionali (unica modalità possibile per specificare i vincoli). Diversamente da quanto visto fino ad ora, i *Dynamic Constraints* permettono di esprimere vincoli *temporali* sui dati attraverso la logica proposizionale, come introdotto da (Chomicki, 1992).

Da un punto di vista implementativo, i vincoli descritti sia da Dynamic Constraints che dalle dipendenze funzionali possono essere realizzati attraverso l'uso di trigger. Alcuni DBMS permettono l'attivazione automatica dei trigger al verificarsi di alcune condizioni (es, il saldo di un conto corrente che scende al di sottodi una soglia stabilita può attivare una procedura interna al sistema). Il potere espressivo dei trigger è tale da esprimere complessi vincoli di data quality, tuttavia, il loro uso è generalmente sconsigliato in casi di grandi database poiché l'attivazione di un trigger può comportare l'attivazione di altri trigger a cascata, generando problemi di attesa circolare (ossia, trigger che si attivano reciprocamente). Non a caso, tecniche di verifica formali (come quelle che verranno descritte nei capitoli successivi) sono state usate per dimostrare formalmente la terminazione di una sequenza di trigger, ovvero per avere la garanzia che un insieme di trigger termini sempre la propria computazione, indipendentemente dai dati contenuti all'interno del database (Ray & Ray, 2001).

Da un punto di vista statistico, invece, è da menzionare la tecnica del *data imputation* (o *data editing*). Le regole di consistenza (chiamate *edits*) vengono definite sulle tabelle. Si assuma di avere una tabella dell'ufficio della Motorizzazione Civile, in cui al codice fiscale è associato il codice della licenza di guida. Una semplice regola di consistenza potrebbe richiedere che la data di nascita derivabile dal codice fiscale garantisca la maggiore età dell'utente alla data di acquisizione della licenza di guida. In caso contrario, una volta individuata una inconsistenza, questa sarà eliminata o candellando l'intero record o *imputando* direttamente un nuovo valore individuato con tecniche statistiche. Un'evoluzione della tecnica fornita da Fellegi e Holt è la New Imputation Methodology (Bankier, 1999), utilizzata principalmente per la gestione dei dati dei censimenti, spesso affetti da inconsistenze e missing item. Un problema ancora aperto nell'uso di questa metodologia è la modalità con la quale identificare il giusto valore di imputazione, storicamente noto in letterature con il nome di *imputation problem* (Fellegi & Holt, 1976).

- **Basati sull'apprendimento:** Se nei metodi basati sulle dipendenze e su business rules è l'esperto di dominio a definire i vincoli (di integrità, di consistenza etc), nell'approccio basato sull'apprendimento è il software che, opportunamente addestrato con dei dataset ideali chiamati "training dataset", individua ed eventualmente corregge il dato sulla base di quanto appreso nella fase di training. La letteratura scientifica propone una grande varietà di

approcci, molti dei quali mutuati dall'Intelligenza Artificiale e dalla Statistica (es., Neural Network, Genetic Algorithms, Pattern Recognition, Clustering Algorithm, etc)**.** La maggiore difficoltà nell'uso delle tecniche di apprendimento risiede nella generazione di un opportuno training dataset, cioè di un dataset ideale capace di trattare ogni tipologia di dato conforme al dataset di training. Generalmente, l'output che questi algoritmi forniscono sono analizzati ulteriormente da esperti di dominio, i quali con meccanismi di rinforzo, possono influire sui successivi cicli di training al fine di migliorare le capacità di apprendimento del software stesso. Una recente ed efficace applicazione di queste tecniche per l'analisi e la messa in qualità dei dati è stata presentata da (Mayfield, Neville, & Prabhakar, 2009).

- **ETL (Extract, Transform, Load):** Il termine ETL identifica un processo che prevede una fase di (1) estrazione dei dati da una o più sorgenti; (2) manipolazione dei dati (che comprende la fase di data quality assesment e cleansing) e (3) caricamento dei dati nel database/data warehouse finale. La fase di data quality è generalmente realizzata con l'ausilio di *business rule* (regole definite dall'esperto di dominio per gestire la fase di messa in qualità dei dati). Una delle difficoltà di questo approccio risiede nella *formalizzazione* delle regole di business così come nel *monitoraggio* dei side-effect (ossia, degli effetti non immediatamente riscontrabili) che l'esecuzione di queste regole può avere sul dato manipolato. A tal proposito, un gran numero di strumenti di *data* profiling sono disponibili sul mercato, i quali permettono di svogere una grande quantità di analisi dei dati in fuzione delle dimensioni prescelte. Una panoramica approfondita sui maggiori strumenti di ETL è raccolta nei riferimenti bibliografici (Galhardas, Florescuand, Simon, & Shasha, 2000), (Batini & Scannapieco, 2006) (Muller & Freytag, 2003).

3. 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 ed analizza il processo di messa in qualità dei dati delle Comunicazioni Obbligatorie, le quali descrivono i principali eventi che caratterizzano l'evoluzione del mercato stesso: avviamenti al lavoro, cessazioni, proroghe di rapporti di lavoro esistenti o loro trasformazioni.

3.1 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/servizi/azienda/argo02/Pagine/default.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 è 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: ciò a portato ad un progressivo miglioramento della qualità del dato raccolto, ma di conseguenza a diversi livelli di qualità dell'informazione in funzione del momento di raccolta del dato e del luogo di raccolta.

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.

3.2 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 sola cessazione riferita ad un rapporto di lavoro 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. Come verrà descritto nel seguito esistono diverse modalità per procedere a tale ricostruzione: in alcuni casi è possibile recuperare l'informazione

direttamente dalla cessazione stessa, in altri è necessario stimare con metodi probabilistici l'istante in cui tale comunicazione dovrebbe essere avvenuta.

Qualche altro esempio può servire a comprendere come i due tipi di informazioni abbiano caratteristiche differenti ed in modo diverso debbano essere trattati. La comunicazione obbligatoria vista come evento amministrativo ha valore puntuale e riguarda solo il presente: viene comunicato l'evento avvenuto in modo che se ne possa tenere traccia negli archivi amministrativi. Dal punto di vista statistico invece, oltre al presente, la comunicazione può avere effetti sia sul passato sia sul futuro: all'interno della comunicazione si possono trovare elementi riferiti al passato (ad esempio l'avviamento a cui si riferisce la cessazione comunicata) o al futuro (ad esempio la data di cessazione prevista). Affinché il dato sia consistente al momento della registrazione della comunicazione, la banca dati statistica deve essere modificata di conseguenza sia nel passato (correggendo o inserendo le informazioni mancanti) sia nel futuro (inserendo la previsione di un evento che potrà poi essere confermato).

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.

3.3 Criticità del processo di messa in qualità

Il processo che verrà descritto di seguito presenta dunque 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 contener effettivamente 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.

3.4 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. Anche le forniture pregresse vengono processate in modo da ricondurre le informazioni al formato utilizzato.

Da un punto di vista tecnico, l'hash è una funzione univoca operante in un solo senso (ossia, che non può essere invertita), atta alla trasformazione di un testo di lunghezza arbitraria in una stringa di lunghezza fissa, relativamente limitata. Tale stringa rappresenta una sorta di "impronta digitale" del testo in chiaro, e viene detta valore di hash, checksum crittografico o message digest. In informatica, la funzione di trasformazione che genera l'hash opera sui bit di un file qualsiasi, restituendo una stringa di bit di lunghezza predefinita. Spesso il nome della funzione di hash include il numero di bit che questa genera: ad esempio, SHA-256 genera una stringa di 256 bit. L'algoritmo di "Hash" elabora qualunque mole di bit. Si tratta di una famiglia di algoritmi che soddisfa i seguenti requisiti:

- L'algoritmo restituisce una stringa di numeri e lettere a partire da un file, di qualsiasi dimensione esso sia. La stringa è detta Digest;
- La stringa è univoca ed identifica univocamente un documento. Perciò, l'algoritmo è utilizzabile per la firma digitale;
- La funzione di hash non è invertibile, ossia non è possibile ricostruire il documento originale a partire dalla stringa che viene restituita in output.

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'algoritmo è stato implementato in linguaggio java in modo da assicurarne la portabilità ai sistemi attualmente più diffusi. È costituito da un eseguibile stand alone e richiede un file di testo in ingresso riportante su ciascuna riga un codice fiscale da anonimizzare e nessun carattere aggiuntivo; l'algoritmo è in grado di processare anche eventuali codici fiscali formalmente non corretti.

A titolo esemplificativo viene riportato il risultato dell'elaborazione di un codice fiscale per mezzo dell'algoritmo di hashing.

Codice fiscale: AAABBB99C88D777E

Corrispondente hashing: 404f58533b241d21fac63671711fa474adabed19 (l'uso dei caratteri minuscoli o maiuscoli è opzionale).

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

3.5 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 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. 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 di norma caratterizzato da una data di inizio, eventualmente da una data di fine 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.
- **Carriera**: i rapporti possono essere ulteriormente aggregati in carriere. Esse rappresentano le successioni temporali di tutti i rapporti instaurati dal medesimo soggetto (lavoratore). Per poter essere analizzate le carriere devono presentare coerenza interna, non devono cioè esistere al loro interno rapporti con periodi di riferimento sovrapposti, salvo che previsti dalla normativa.
- **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.
- **Stock**: lo stock rappresenta un ulteriore livello di aggregazione dei rapporti che però anziché diminuirne la numerosità associando alcuni dei suoi elementi, li

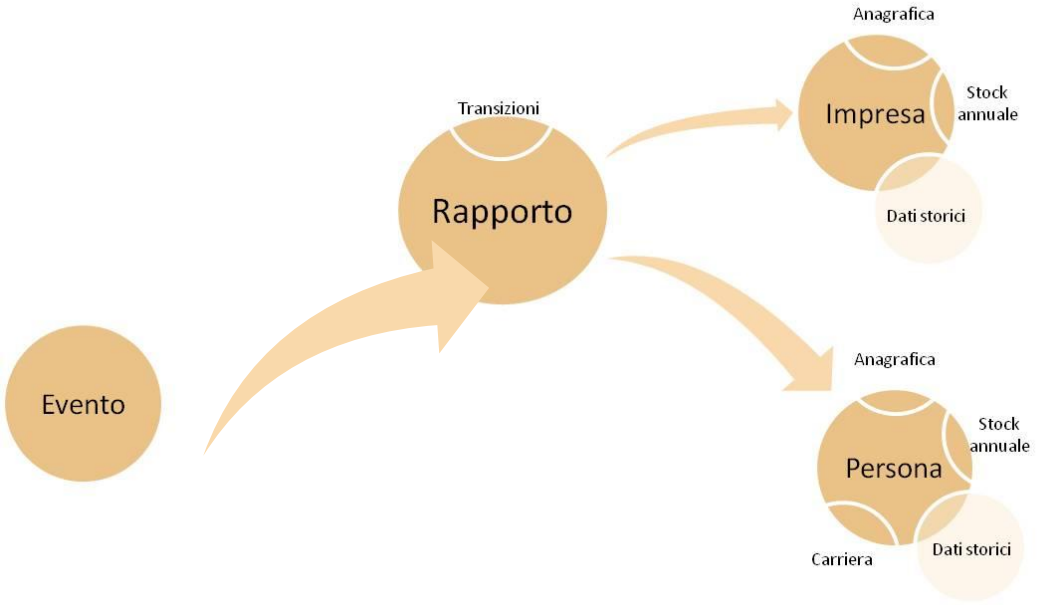
seleziona in base al periodo temporale di interesse: fanno parte del medesimo stock tutti i rapporti che si svolgono, anche solo in parte, all'interno del periodo (l'anno in prima analisi) di interesse.

- **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, i dati reddituali, 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.
- **Livelli di aggregazione:** una componente fondamentale per la corretta comprensione dell'intero modello è rappresentata dal concetto di aggregazione. L'aggregazione può avvenire all'interno del modello secondo due modalità: *aggregazione logica*, che porta ad associare in un unico elemento singole entità aventi caratteristiche comuni (ad esempio l'aggregazione dei rapporti in carriere) e *l'aggregazione temporale*, che porta a relazionare fra di loro entità aventi lo stesso periodo temporale di interesse. Data la natura delle fonti utilizzate è possibile che uno stesso elemento (la persona ad esempio) sia caratterizzata da elementi associati a livello logico (la carriera) ed altri a livello temporale (gli stock su base annuale). È necessario porre particolare attenzione nel mettere in relazione tra loro elementi a livelli di aggregazione differenti pur associati alla medesima entità.
- **Relazioni**: le relazioni sono l'elemento base su cui si poggia l'intero modello: la relazione tra evento, persona ed impresa costituisce il fulcro fondamentale dell'analisi, ma è arricchita da diverse altre relazioni come quelle esistenti tra rapporto e rapporto e tra eventi e rapporti, solo per citare alcuni esempi.

Per agevolare la comprensione del modello di riferimento complessivo viene di seguito presentato il processo logico che porta alla sue costruzione.

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.



A partire dai rapporti possono essere generate le prime due entità aggregando gli elementi in base a criteri temporali: 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 i rapporti 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: la carriera contiene ad esempio attributi caratterizzanti l'intera filiera dei rapporti riguardanti la persona considerati nel loro complesso; 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 i rapporti 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.

3.6 Il processo di messa in qualità

Dal punto di vista operativo il processo di messa in qualità delle comunicazioni obbligatorie avviene attraverso il popolamento di tre diverse strutture:

- Il repository delle fonti: contiene le sorgenti informative nel loro formati originale e rappresenta anche lo storico da cui ripartire in caso di eventuali rielaborazioni;
- L'area di staging: una volta caricato in questa area il dato può essere modificato e messo in qualità in funzione delle regole identificate. Nel corso del trattamento il dato viene anche aggregato e vengono create all'interno dell'area le strutture di riferimento (ad esempio per i rapporti). Il dato in questa area viene trattato in forma normalizzata;
- L'area di analisi: contiene il dato al termine della messa in qualità, pronto per essere analizzato. Il dato in questa area viene trattato in forma denormalizzata.

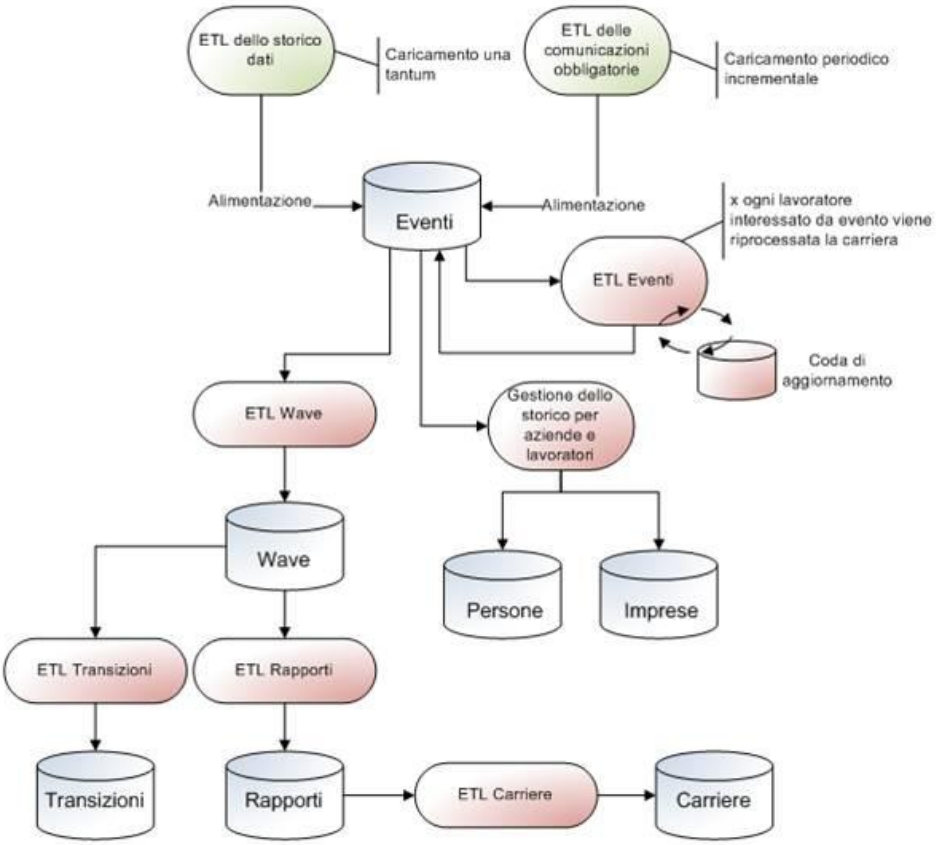
Nel seguito del documento verranno descritti i processi operati all'interno dell'area di staging.

4. Descrizione del Processo di ETL

4.1 Modello di trattamento complessivo

Nel presente paragrafo viene illustrato il modello di trattamento complessivo a partire dalle sorgenti dati fino alle generazione delle tabelle finali di staging, dalle quali verranno poi generate le tabelle di analisi. Nel corso del processo vengono dapprima caricati i dati provenienti dalle sorgenti quindi vengono rielaborati gli eventi: per garantire la coerenza della successione dei nuovi eventi con gli eventi passati, in questa fase di elaborazione vengono ricaricati anche tutti gli eventi corrispondenti ai soggetti riferiti all'interno delle comunicazioni caricate. Come si vedrà nel seguito infatti il caricamento di un nuovo evento all'interno della carriera di un lavoratore comporta il trattamento ex novo di tutta la carriera nel suo complesso. L'elaborazione non riguarda quindi solamente le nuove comunicazioni, ma anche tutte le comunicazioni riferite allo stesso soggetto e già presenti nel database.

Le sorgenti che concorrono a definire gli eventi principali che caratterizzano il modello dati sono come descritto in precedenza le Comunicazioni Obbligatorie. Le comunicazioni obbligatorie possono essere ulteriormente distinte tra comunicazioni storiche (antecedenti all'anno 2008 cioè all'entrata in vigore dell'obbligatorietà delle comunicazioni obbligatorie telematiche) e comunicazioni obbligatorie telematiche; le due tipologie si distinguono per i contenuti informativi, per le modalità di trasmissione e per le classificazioni utilizzate e devono quindi a tutti gli effetti essere considerate come sorgenti informative distinte pur concorrendo al popolamento della medesima tipologia di eventi. Inoltre le comunicazioni obbligatorie telematiche presentano un formato definito univocamente a livello nazionale e non richiedono dunque personalizzazioni a seconda del territorio in esame. Le comunicazioni storiche invece, strettamente legate alla modalità di registrazione in uso nel territorio in esame, possono presentare a seconda dei casi sostanziali differenze e richiedere dunque forti personalizzazioni del processo di caricamento.



La disponibilità delle informazioni a partire dalle sorgenti può avvenire secondo diverse modalità:

- Una tantum: è il caso ad esempio delle comunicazioni obbligatorie storiche, riguardanti un periodo passato, e che vengono caricate una sola volta all'interno del sistema al momento dell'inizializzazione;
- Modalità non periodica: è il caso delle forniture per cui non è prevedibile il momento in cui saranno disponibili e il cui trattamento può dunque essere attivato nel momento stesso in cui i dati vengono forniti;
- Modalità periodica: è il caso delle forniture disponibili ad intervalli di tempo regolari. Il relativo trattamento può dunque essere attivato con programmazione regolare con cadenza definita in base alle esigenze degli utenti. Il processo di trattamento è comunque progettato indipendentemente rispetto alla periodicità indicata.

Il processo prevede meccanismi di caricamento incrementali: l'arrivo di nuove comunicazioni non prevede la rielaborazione dell'intera banca dati, ma nemmeno per quanto descritto in precedenza la semplice elaborazione dei nuovi casi. Ogni aggiornamento prevede la rielaborazione di tutti gli eventi (nuovi e non) riferiti ai soggetti indicati nelle nuove comunicazioni pervenute.

Una volta completata la fase di estrazione e pretrattamento degli eventi è possibile avviare la fase di trattamento vera e propria che porta, attraverso una serie di passaggi successivi, al popolamento delle diverse strutture di analisi. Tale fase viene progettata seguendo un approccio incrementale: ciascun record in ingresso viene trattato singolarmente e integrato con i dati esistenti. Tale modalità consente di diminuire i tempi di elaborazione (solo le informazioni interessate vengono trattate, modificate e salvate nel corso di ciascun aggiornamento) e di svincolare la periodicità del trattamento dalla periodicità delle estrazioni a partire dalle fonti, non necessariamente sincrone. Per consentire tale modalità e per ottimizzare ulteriormente il processo di trattamento viene mantenuta in ingresso una coda di soggetti da sottoporre ad aggiornamento, con la struttura di una coda FIFO (first in first out): in tal modo anche in presenza di più record riguardanti il medesimo soggetto il trattamento viene eseguito una sola volta prelevando tutti i nuovi record che lo riguardano e modificandone in maniera omogenea la carriera. La coda in ingresso viene svuotata man mano che si procede al caricamento delle informazioni. Per ciascun soggetto presente nella coda in ingresso viene ripetuto l'intero processo di caricamento perché tutte le strutture possono essere interessate dalle nuove informazioni e queste ultime devono essere rese omogenee rispetto a tutte le informazioni già presenti. Nel seguito verranno descritte in maggiore dettaglio le principali operazioni svolte al suo interno.

Una volta completato il processo di trattamento degli eventi, il risultato di tale elaborazione diviene a sua volta sorgente per:

- Il processo di aggiornamento delle anagrafiche (persone ed imprese) aggiornando dove necessario il dato riferito ai soggetti trattati o inserendo i soggetti mancanti;
- Il processo di generazione dei rapporti, eseguito aggregando gli eventi riferiti allo stesso rapporto. Quest'ultimo processo è preceduto da un passaggio preliminare di generazione della wave, uno stadio intermedio tra eventi e rapporti che considera come differenti rapporti aventi anche una sola carattersitica discordante (ad esemio la qualifica), ed alimenta anche il processo di generazione delle transizioni, dei passaggi cioè tra diversi rapporti (o diverse wave);
- Il processo di generazione delle carriere ottenute come aggregazione di rapporti riferiti al medesimo soggetto.

Nel presente documento non vengono descritti tutti i passaggi ma viene posta l'attenzione su:

- L'estrazione dei dati a partire dalle fonti informative coinvolte;
- Il trattamento preliminare dei dati sorgente;
- La riconduzione delle informazioni ad un modello unificato e la relativa transcodifica delle classificazioni utilizzate;
- Il trattamento dei dati e la loro messa in qualità.

Queste attività sono infatti maggiormente interessate dal processo di messa in qualità ed è al loro interno che vengono applicate le regole di correzione e generazione che contribuiscono al miglioramento della qualità dei dati.

4.1.1 Caricamento del dato storico proveniente dai SIL

Il componente di caricamento del dato storico prevede la riconduzione delle informazioni provenienti dai sistemi informativi per il lavoro delle diverse realtà territoriali antecedenti al 2008 al modello dati unificato degli eventi utilizzando le classificazioni standard.

- Caricamento dati: il caricamento dei dati avviene a partire da sorgenti flat file o da una banca dati sorgente per mezzo di un caricamento diretto. Eventuali aggregazioni o integrazioni di informazioni presenti separatamente sulla fonte originale avvengono in questa fase. Al termine di questa fase i dati sono presenti sul sistema destinatario in memoria pronti ad essere trattati.
- Applicazione algoritmi di cripting: per garantire l'anonimato dei soggetti analizzati tutte le informazioni sensibili che consento di ricondurre i dati ad un particolare soggetto (codice fiscale, partita IVA, ecc) vengono criptate per mezzo di un algoritmo di hashing irreversibile che consente di mantenere l'univocità del codice pur non consentendo l'identificazione del soggetto. Tutte le sorgenti caricate vengono sottoposte al medesimo trattamento di hashing in modo da consentire nei passaggi successivi l'integrazione delle informazioni riguardanti lo stesso soggetto pur non essendo a conoscenza della sua identità. Si noti che ulteriori informazioni che consentono l'identificazione del soggetto (nome, cognome, ecc..) non vengono caricate non essendo di interesse ai fini delle analisi.
- Deduplica logica: la deduplica logica comporta l'individuazione e l'aggregazione di record che pur presentando identificativi fisici differenti sono analoghi dal punto di vista logico. A tal fine è necessario individuare oltre alle chiavi fisiche alcune chiavi logiche la cui identità comporta l'analogia logica. Per quanto riguarda le comunicazioni obbligatorie la chiave logica è rappresentata dalla data dell'evento, dal lavoratore interessato, dalla sede operativa dell'impresa e dal tipo di comunicazione. Ad esempio due eventi identificati da codici diversi (provenienti quindi da due comunicazioni diverse) ma aventi i campi indicati in precedenza uguali vengono considerate come un'unica comunicazione.
- Trasformazione delle comunicazioni in singoli eventi: è possibile che un singolo record del sistema sorgente contenga informazioni riguardanti più eventi. Per consentire la corretta associazione dei dati agli eventi in questa fase tali record vengono suddivisi in più record ciascuno riportante l'informazione collegata ad un singolo evento. Ad esempio un record potrebbe contenere informazioni riguardanti sia un avviamento sia la sua prima trasformazione: in tal caso esso viene suddiviso in due record, ciascuno dei quali riporta le informazioni di uno dei due eventi. Oppure una singola comunicazione potrebbe contenere sia la data di avviamento sia la data di cessazione prevista (nel caso di avviamenti a tempo determinato): vengono quindi generate due comunicazioni di cui una collocata nel futuro rispetto allla comunicazione caricata e riportante l'evento di cessazione previsto.
- Transcodifica: ciascun sistema sorgente adotta classificazioni proprietarie. Per consentire l'integrazione dei dati è necessario che le classificazioni vengano ricondotte agli standard adottati. Tale passaggio avviene per mezzo di tabelle di transcodifica che associano ciascuno dei valori utilizzati nel sistema sorgente al corrispondente valore nella classificazione standard. Nel caso tale associazione non esista sono previste voci generiche a cui ricondurre l'informazione; nel caso

l'informazione sia mancante sono previste apposite voci. Ad esempio la classificazione delle qualifiche viene aggiornata periodicamente: ad ogni aggiornamento tutte le voci precedentemente in uso vengono ricodificate e ricondotte alla nuova classificazione. Ciò consente a tutta la banca di tuilizzare le medesime codifiche indipendentemente dal momento di caricamento o dall'istante di osservazione.

- Mapping: il passaggio conclusivo di questa fase comporta la riconduzione del modello dati sorgente al modello dati degli eventi, operata riconducendo ciascun campo al campo corrispondente nel nuovo modello. Eventuali informazioni non utilizzate nel nuovo modello vengono scartate. Eventuali informazioni mancanti vengono valorizzate con apposite costanti. Ad esempio informazioni non utilizzate nel corso delle analisi vengono escluse (si pensi ai riferimenti normativi legati al trattamento previdenziale) , eventuali campi aggiunti nel corso del tempo alle comunicazioni obbligatorie vengono valorizzati solo dove presenti e campi che hanno cambiato nome nel corso degli anni vengono ricondotti ad un unico standard.
- Filtri temporali: il consolidarsi della serie storica fornita dalle comunicazioni obbligatorie telematiche ha reso man mano meno importante il contributo dei dati provenienti dallo storico conservato all'interno dei SIL locali. È possibile impostare un filtro temporale per caricare non tutte le comunicazioni presenti nei SIL locali, ma solo quelle a partire da una certa data, avendo cura di importare anche eventi precedenti ma con impatto successivo a quella data (ad esempio un avviamento precedente alla data di filtro ma ancora aperto al momento della data limite). Attualmente tale filtro è impostato al 1 gennaio 2007.

4.1.2 Caricamento comunicazioni obbligatorie telematiche

Il componente di caricamento delle comunicazioni obbligatorie telematiche prevede la riconduzione delle informazioni provenienti dai sistemi informativi per il lavoro delle diverse realtà territoriali o dal sistema nazionale a partire dall'anno 2008 al modello dati unificato degli eventi utilizzando le classificazioni standard.

- Parsing delle comunicazioni obbligatorie: prima di avviare il caricamento è necessario ricondurre il formato xml originale ad un formato flat utilizzabile dai processi successivi. Tale processo viene eseguito per mezzo di un parser (cioè di uno strumento informatico che processa la comunicazione ed estrae i singoli campi di interesse, riconoscendoli all'interno del file xml e posizionandoli correttamente all'interno del tracciato record) che processa ogni singolo file xml e ne estrae le informazioni di interesse all'interno di un file successivamente utilizzato in ingresso ai passaggi seguenti. Tale componente effettua una prima selezione delle informazioni escludendo dalla fornitura i record non corrispondenti al tracciato originale. Per ciascuna tipologia di comunicazione è previsto un apposito parser poiché ciascuna di esse presenta un tracciato xml e contenuti differenti. Il componente si occupa di distinguere tra le diverse tipologie di comunicazione e di applicare ad esse il corretto parser.
- Caricamento dati: il caricamento dei dati avviene a partire dai file prodotti dal sotto componente precedente. Al termine di questa fase i dati sono presenti sul sistema destinatario in memoria pronti ad essere trattati.

- Applicazione algoritmi di cripting: per garantire l'anonimato dei soggetti analizzati tutte le informazioni sensibili che consento di ricondurre i dati ad un particolare soggetto (codice fiscale, partita IVA, ecc) vengono criptati per mezzo di un algoritmo di hashing irreversibile che consente di mantenere l'univocità del codice per non consentendo l'identificazione del soggetto.
- Deduplica logica: la deduplica logica comporta l'individuazione e l'aggregazione di record che pur presentando identificativi fisici differenti sono analoghi dal punto di vista logico. A tal fine è necessario individuare oltre alle chiavi fisiche alcune chiavi logiche la cui identità comporta l'analogia logica. Per quanto riguarda le comunicazioni obbligatorie la chiave logica è rappresentata dalla data dell'evento, dal lavoratore interessato, dalla sede operativa dell'impresa e dal tipo di comunicazione. Ad esempio due eventi identificate da codici diversi (provenienti quindi da due comunicazioni diverse) ma aventi i campi indicati in precedenza uguali vengono considerate come un'unica comunicazione.
- Gestione degli annullamenti e delle rettifiche: le comunicazioni obbligatorie oltre a comunicare un evento possono comunicare anche rettifiche o annullamenti di comunicazioni precedenti. Tali comunicazioni devono essere gestite in fase di caricamento per garantire la consistenza del dato trattato. Tali comunicazioni "speciali" rientrano nelle seguenti categorie:
  - Annullamenti: la comunicazione annulla una comunicazione inviata in precenza. La comunicazione riferita viene cercata e, se individuata, eliminata dal set di elaborazione. La comunicazione di annullamento viene poi eliminata dal set di elaborazione. Non è sempre possibile rintracciare la comunicazione riferita da cui deriva la differenza tra comunicazioni annullate e comunicazione di annullamento.
  - Rettifiche: la comunicazione corregge una comunicazione precedente. La comunicazione riferita viene cercata e, se indivduata, eliminata dal set di elaborazione. La comunicazione di rettifica resta a far parte del set di elaborazione al posto della precedente. Non è sempre possibile rintracciare la comunicazione riferita da cui deriva la differenza tra comunicazioni rettificate e comunicazione di rettifica.
  - Comunicazioni a seguito d'urgenza: la comunicazione entra a far parte del set di elaborazione. Sarà seguita da una comunicazione standard: in tal caso il processo di deduplica logica provvederà a risolvere la sovrapposizione.
  - Trasformazione da tirocinio in rapporto di lavoro subordinato: la comunicazione viene gestita come una normale trasformazione ed inserita nel set di elaborazione.
  - Inserimenti d'ufficio: la comunicazione, pur se inserita manualmente, viene gestita come una normale trasformazione ed inserita nel set di elaborazione.

La tabella riportata di seguito mostra come la gestione di tali comunicazioni sia fondamentale: pur rappresentando solo una parte delle comunicazioni complessive la loro numerosità nel tempo genera modifiche significative alla banca dati. Le numerosità riportate in tabella sono riferite alle comunicazioni obbligatorie della regione Lombardia fino al Settembre 2012.

Tipo di comunicazione	Numerosità
Comunicazioni standard	15.877.471
Annullamenti	222.615
Annullate	212.104
Rettificate	494.388
Rettifiche	526.348
Comunicazioni a seguito d'urgenza	26.551
Trasformazione da tirocinio in rapporto di lavoro subordinato	6.138
Inserimenti d'ufficio	90.908
Comunicazioni in ingresso	16.750.031
Comunicazioni caricate dopo la gestione	15.773.601

- Trasformazione delle comunicazioni obbligatorie telematiche in singoli eventi: è possibile che un singolo record del database sorgente contenga informazioni riguardanti più eventi. Per consentire la corretta associazione dei dati agli eventi, in questa fase tali record vengono suddivisi in più record ciascuno riportante l'informazione di un singolo evento.
- Transcodifica: ciascun sistema sorgente adotta classificazioni proprietarie. Per consentire l'integrazione dei dati è necessario che le classificazioni vengano ricondotte agli standard adottati. Tale passaggio avviene per mezzo di tabelle di transcodifica che associano ciascuno dei valori utilizzati nel sistema sorgente al corrispondente valore nella classificazione standard. Nel caso tale associazione non esista sono previste voci generiche a cui ricondurre l'informazione; nel caso l'informazione sia mancante sono previste apposite voci.
- Mapping: il passaggio conclusivo di questa fase comporta la riconduzione del modello dati sorgente al modello dati degli eventi, operata riconducendo ciascun campo al campo corrispondente nel nuovo modello. Eventuali informazioni non utilizzate nel nuovo modello vengono scartate. Eventuali informazioni mancanti vengono valorizzate con apposite costanti.

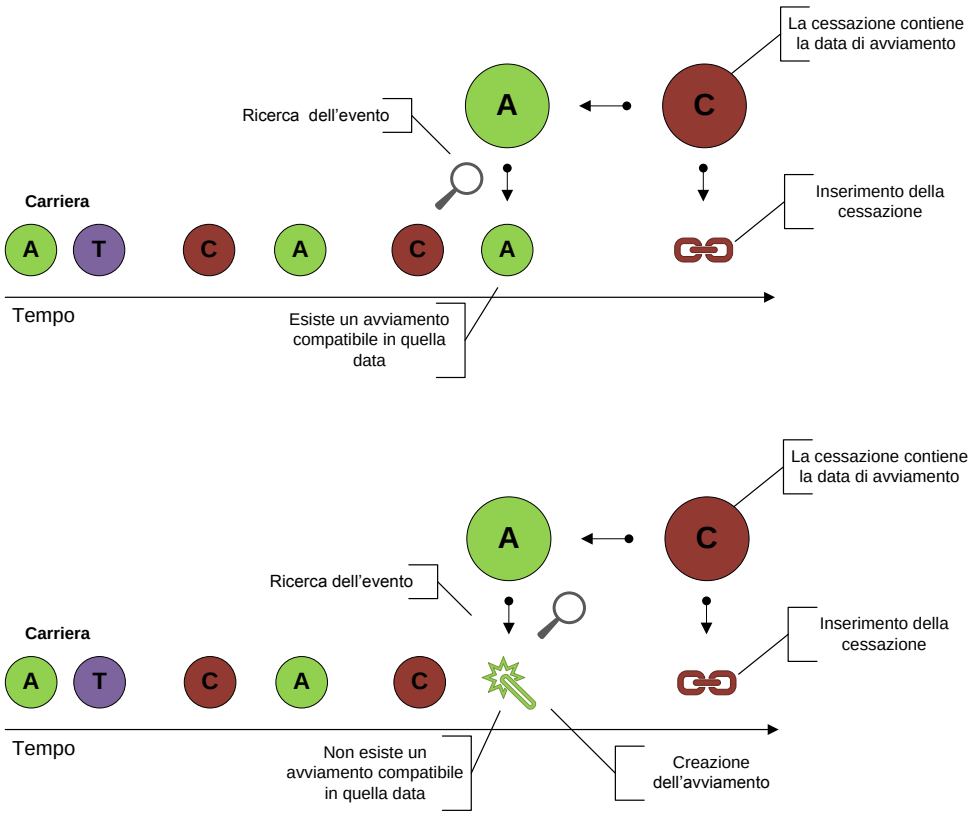
4.1.3 Trattamento degli eventi

Il trattamento degli eventi ha l'obiettivo di ricondurre gli eventi riguardanti il medesimo soggetto al modello dati utilizzato, metterli in qualità e renderli omogenei tra di loro. Questa fase rappresenta il nucleo centrale del processo di messa in qualità. Questa primo paragrafo ne descrive separatamente le componenti logiche. Per consentire la comprensione dei meccanismi che operano all'interno di questa attività, nei due paragrafi successivi verranno presentati il modello logico alla base di questa fase, alcuni esempi di trattamento di carriere e una breve sintesi delle numerosità coinvolte nel processo descritto.

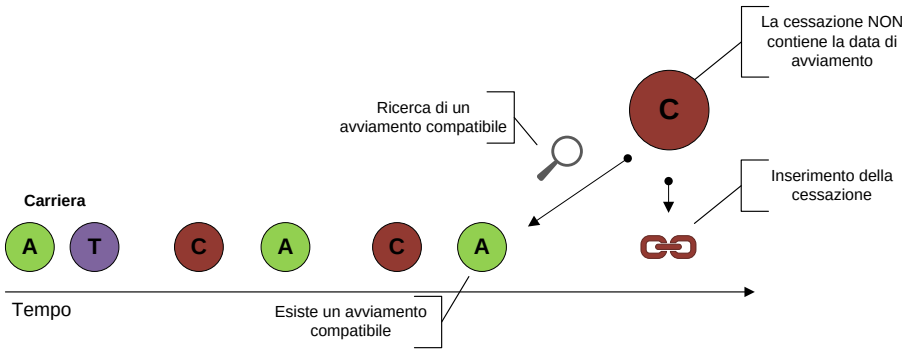
- Deduplica degli eventi: questo passaggio, condotto all'interno di ogni singola fornitura, mira ad individuare record distinti facenti riferimento ad un singolo evento e a fonderli in un unico record. A tal fine è necessario individuare oltre alle chiavi fisiche alcune chiavi logiche la cui identità comporta l'analogia logica. Sostanzialmente rappresenta una ripetizione dei passaggi di deduplica precedente operati non più sulla singola fornitura ma sull'unione delle diverse forniture.

L'operazione di deduplica viene eseguita più volte nel corso del processo di messa in qualità per garantire che anche a fronte di modifiche o generazioni di eventi venga mantenuta l'univocità degli stessi.

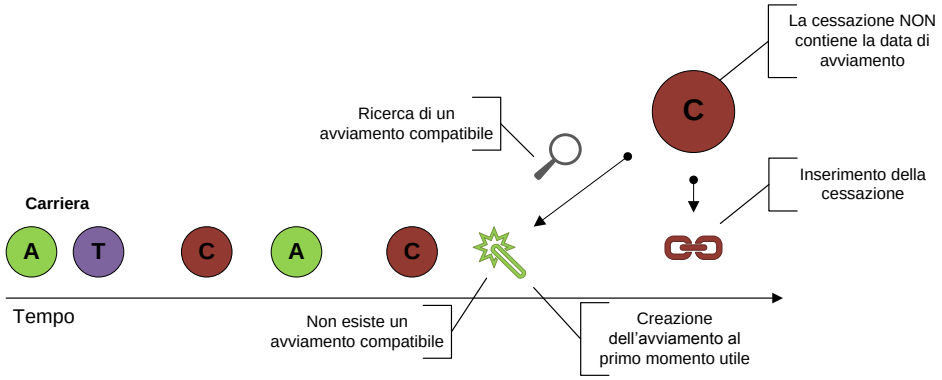
- Ricostruzione degli avviamenti mancanti: questo passaggio garantisce la coerenza della successione degli eventi all'interno di una carriera assicurando che una volta registrata una comunicazione di cessazione sia sempre presente il corrispondente evento di avviamento. Se la cessazione riporta al suo interno anche la data di avviamento, l'evento viene ricercato tra gli eventi già registrati: se presente viene confermato, altrimenti viene creato un evento di avviamento corrispondente alle caratteristiche individuate nella cessazione.

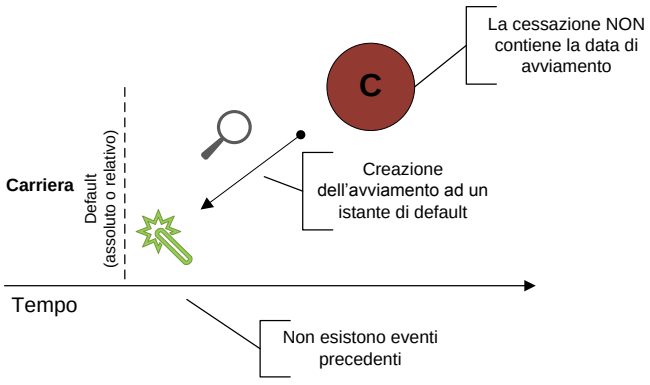


Se invece la cessazione non riporta informazioni riguardanti l'avviamento, viene cercato all'interno dell'archivio un evento di avviamento compatibile (stesso soggetto, stessa azienda, ecc.): se l'evento viene inidivduato viene assunto come avviamento relativo alla cessazione introdotta. La ricerca oltre che agli avviamenti viene estesa anche a proroghe e trasformazioni in quanto potrebbero anch'esse riportare i dati necessari per la correlazione alla cessazione.



Se infine non viene individuato alcun evento compatibile con la cessazione analizzata viene generato ex novo un avviamento: in caso la carriera presenti eventi precedenti a quello analizzato l'avviamento viene generato in corrispondenza del primo periodo "utile" (un periodo in cui cioè non siano presenti altri rapporti di lavoro); nel caso in cui invece non siano presenti eventi precenti l'avviamento può essere posizione in un istante di default, fisso o variabile in base alla data di cessazione: è possibile cioè definire la collazione di default in modo assoluto (ad es. il 1 gennaio 2007) o in modo relativo (x giorni prima della cessazione).

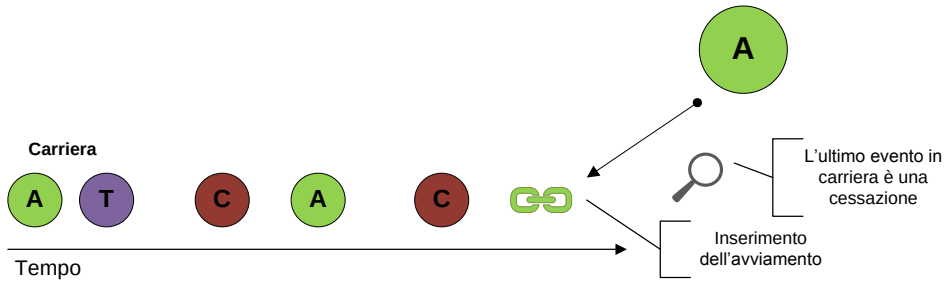




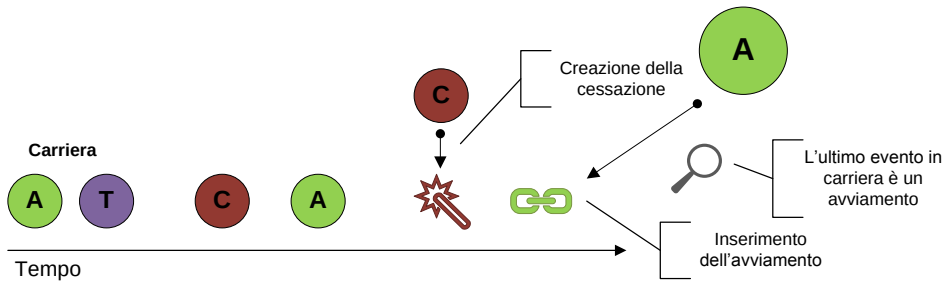
Oltre che nel caso di caricamento di una nuova cessazione la stessa logica viene applicata anche nel caso di caricamento di una proroga o trasformazione. In particolare in questi casi in mancanza di informazioni riguardanti l'avviamento del rapporto prorogato o trasformato, viene generato un avviamento al posto della proroga o della trasformazione stessa.

Il tema del posizionamento degli eventi generati all'interno della carriera riguarda il problema della sensistività che verrà affrontato nel seguito del documento; in ogni caso le procedure utilizzate sono parametriche e i valori e le logiche utilizzate possono essere modificate in base alle esigenze o al variare del contesto di applicazione.

Ricostruzione delle cessazioni mancanti: analogamente a quanto avviene per gli avviamenti anche le cessazioni dove necessario possono essere generate. A fronte della comunicazione di un nuovo avviamento viene verificato lo stato del soggetto interessato: se l'evento precedente riguardante il soggetto è una cessazione, l'avviamento può essere inserito senza problemi.



Se invece l'evento precedente è un avviamento (o una proroga o una trasformazione) il rapporto in essere deve essere chiuso prima di poterne generare uno nuovo: viene pertanto creata la cessazione del rapporto precedente per poi procedere al caricamento del nuovo avviamento.

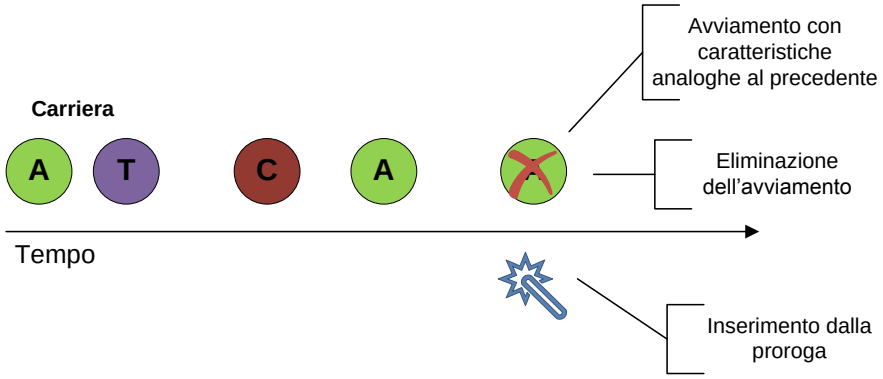


Anche in questo caso il posizionamento della cessazione del tempo coinvolge il tema della sensitività: la cessazione può essere ad esempio inserita nel primo giorno disponibile precedente all'avviamento, ma la logica applicata nel corso del processo è configurabile in base alle esigenze.

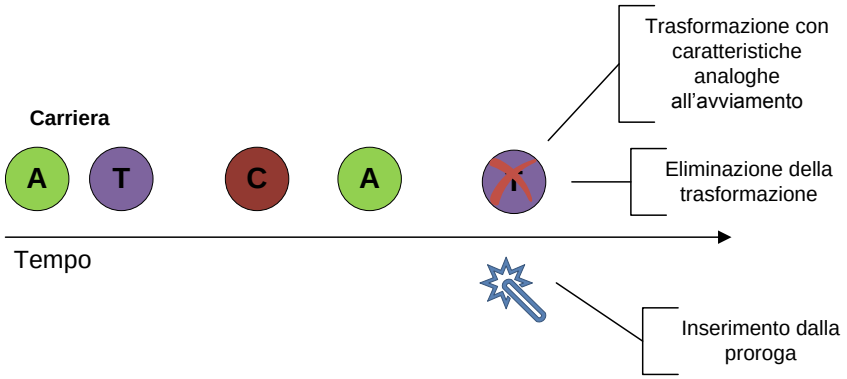
Quanto detto vale naturalmente per i rapporti a tempo pieno che non possono prevedere sovrapposizione: il caso dei rapporti part time o dei rapporti che più genericamente possono precedere sovrapposizoni verrà descritto nel seguito.

- Riclassificazione delle proroghe: le proroghe rappresentano il prolungamento di un contratto di lavoro in essere oltre il periodo di durata previsto. Il processo di trattamento tenta di riclassificare tutti gli eventi successivi ad un avviamento come proghe se essi mantengono inalterate le carattersitche del rapporto ad eccezione della data di termine del rapporto e se sono compatibili con il concetto di proroga.

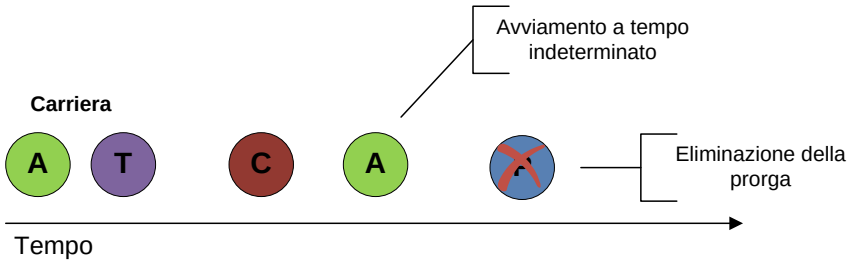
Ad esempio un avviamento che segue un altro avviamento spostandone la data di termine prevista viene riclassificato come proroga.



Analogamente una trasformazione che segue un avviamento ma non ne modifica le caratteristiche ad eccezione della data di termine viene riclassificato come proroga.

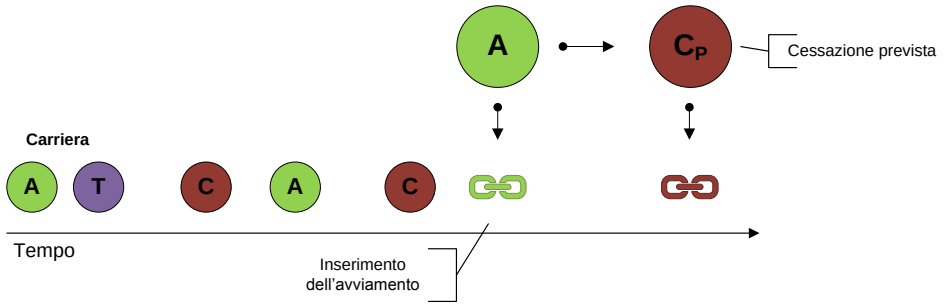


Invece una proroga riferita ad un rapporto a tempo indeterminato non viene considerata valida e viene eliminata.

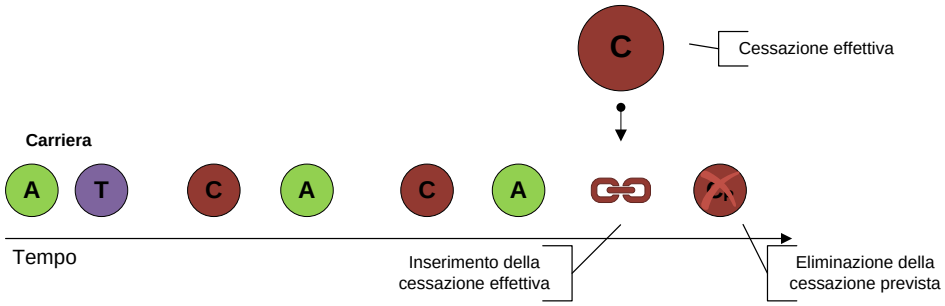


- Riclassificazione delle trasformazioni: le trasformazioni vengono gestite all'interno del processo in maniera analoga a quanto esposto in precedenza per le proroghe ad eccezione del fatto che la trasformazione per essere definita tale deve modificare uno dei parametri del contratto (tipo di contratto, modalità di lavoro ecc.). Qualora una comunicazioni modifichi sia i parametri del contratto (trasformazione) sia i suoi termini temporali (proroga) essa viene classificata come trasformazione.
- Gestione delle cessazioni previste: le cessazioni, oltre ad essere comunicate esplicitamente, possono anche essere incluse implicitamente nella comunicazione di avviamento nel caso di rapporti a termine. In tal caso i dati della cessazione vengono comunicati contestualmente all'avviamento. È dunque necessario che la cessazione venga caricata e gestita adeguatamente, anche nel caso in cui alla comunicazioni implicita faccia seguito quella esplicita. Al momento del

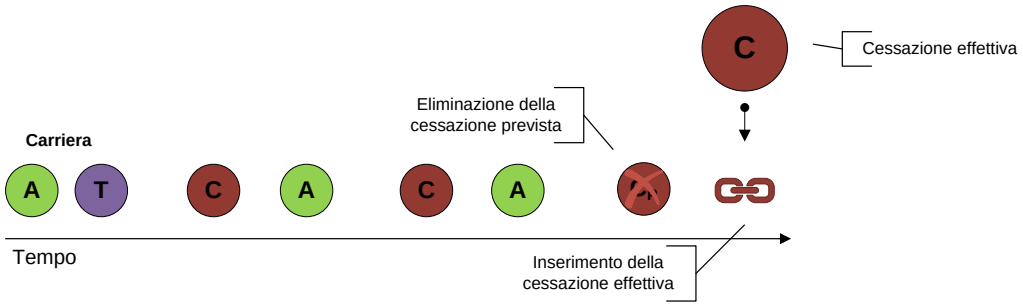
caricamento di un avviamento a termine che riporti anche i dati della cessazione viene generato un evento di cessazione definito come "cessazione prevista". Se nel frattempo non intervengono ulteriori comunicazioni la cessazione diviene definitiva al momento del raggiungimento della data prevista.



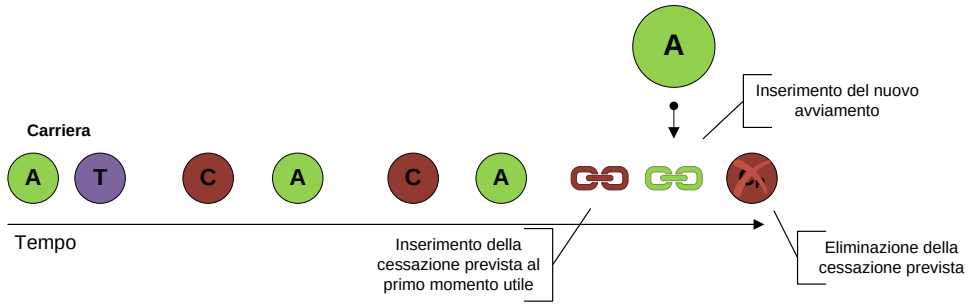
Se invece nel frattempo viene ricevuta una comunicazione esplicita precedente alla data prevista, questa cessazione viene assunta come valida e quella prevista viene eliminata.



Anche nel caso in cui la cessazione comunicata sia successiva a quella prevista, la prima viene assunta come valida poiché comunicata esplicitamente mentre la cessazione prevista viene rimossa.

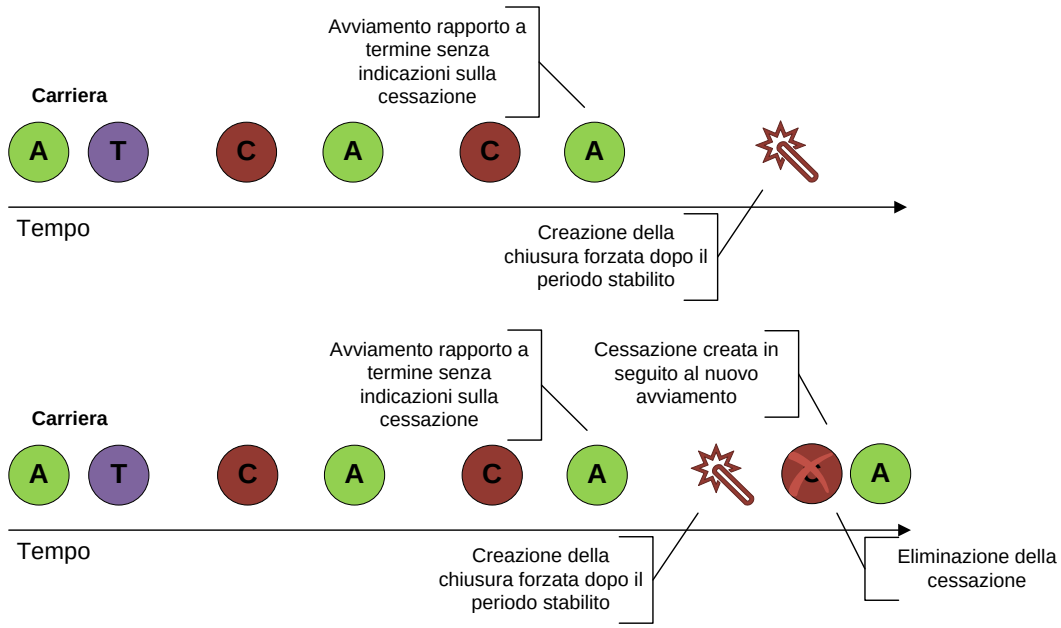


Naturalmente anche le cessazioni previste vengono sottoposte a tutte le logiche di messa in qualità descritte in precedenza per le cessazioni: se ad esempio prima della cessazione prevista viene registrato un nuovo avviamento la cessazione viene spostata all'ultimo giorno utile prima del nuovo avviamento.



La gestione delle cessazioni descritta introduce naturalmente eventi anche nel futuro rispetto al momento di caricamento. Tali eventi, non distinguibili da quelli consolidati, devono essere opportunamente trattati in fase di analisi.

- Chiusura prestabilita di particolari tipologie contrattuali: per evitare che in banca dati si verifichino casi in cui i rapporti legati ad alcune tipologie contrattuali abbiano una durata eccessiva, è stata introdotta una chiusura automatica. Può infatti accadere che venga comunicato l'avviamento di un rapporto a termine, senza poi ricevere più alcuna comunicazione di cessazione. Se il rapporto non è seguito da comunicazioni riguardanti rapporti successivi esso resterebbe aperto.



Le tipologie contrattuali interessate da questa verifica sono:

- Gli apprendistati;
- I contratti a progetto;
- I contratti di lavoro somministrato;

Si è scelto di non prendere in considerazione le altre due principali tipologie contrattuali presenti in banca dati per i seguenti motivi:

- Tempo indeterminato: tale tipologia contrattuale non costituisce elemento di valutazione data la sua caratteristica di contratto non a termine;
- Tempo determinato: con l'avvento delle comunicazioni obbligatorie telematiche, per i rapporti a tempo determinato deve essere già comunicata obbligatoriamente la data di fine rapporto prevista. A seguito di quest'obbligo all'interno del nostro sistema di trattamento dati viene già creata di default la cessazione del rapporto lavorativo.
  Si è verificato che solo lo 0,5% delle comunicazioni a tempo determinato sulla totalità delle comunicazioni non riporta né la data di cessazione, né la data di cessazione prevista, né la data di fine proroga (equivalente alla data di cessazione). Con un ulteriore controllo si è constatato che le comunicazioni che non hanno nessuno di questi campi valorizzati sono tutte trasformazioni e la maggior parte di esse comunica un passaggio dalla tipologia contrattuale a tempo determinato a quella a tempo indeterminato.

La soluzione metodologica utilizzata per la chiusura forzata dei rapporti consiste nell'individuare un limite di durata contrattuale massima ed assegnare tale valore ai rapporti rimasti aperti o le cui durate sono maggiori di tale valore. Per l'individuazione della soglia di durata si è scelto di utilizzare un criterio basato sul range interquartile. Siano Q1 e Q3 il primo e il terzo quartile della distribuzione delle durate contrattuali per ciascuna tipologia, si definisce outlier una osservazione il cui valore misurato è fuori dal range:

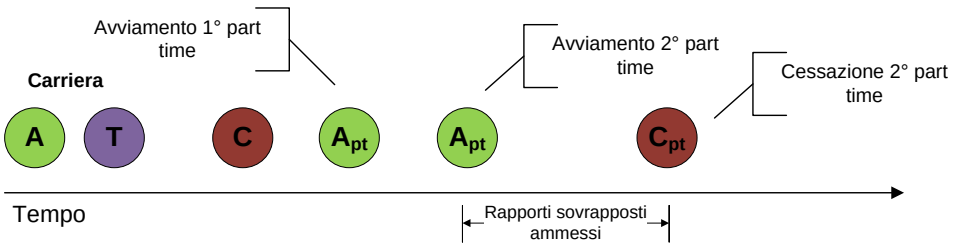
[Q1-k(Q3-Q1);Q3+k(Q3-Q1)] per un prefissato valore k.

Tipicamente k viene posto uguale a 1.5 e le osservazioni fuori dalla soglia vengono definite outliers deboli, oppure può essere posto uguale a 3, si definiscono così gli outliers forti. In questo è stata ritenuta migliore la scelta di k uguale a 1,5. Di seguito vengono illustrate le proposte di chiusura attualmente associate ai contratti presi in considerazione.

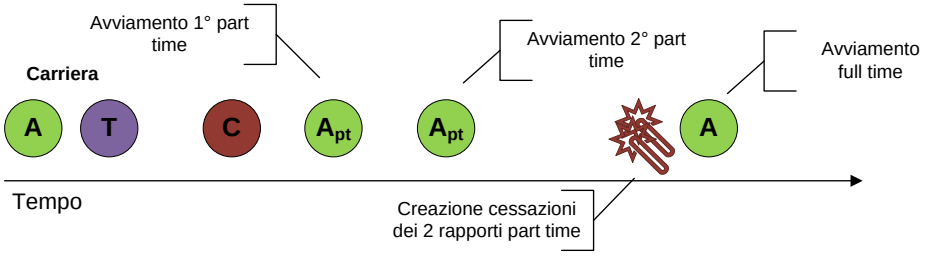
- Apprendistato: il valore risultante secondo il criterio adottato per l'individuazione della soglia è di 2.036 giorni equivalente a circa 5 anni e mezzo. Essendo il valore molto elevato, è stata effettuata una ricerca per verificare se esista un limite di legge per i contratti di apprendistato: per la tipologia contrattuale dell'apprendistato esiste una regolamentazione che fissa la durata massima di un contratto a 4 anni e, solo in alcuni casi particolari, la durata può estendersi fino ai 6 anni. Inoltre la distribuzione per quartili della durata degli apprendistati mostra come il 75% dei contratti raggiunga una durata di poco superiore ai 2 anni. Facendo seguito a tali considerazioni è stata adottata come soglia di chiusura il valore di 1.461 giorni corrispondente a 4 anni.
- Lavoro interinale: la metodologia proposta assegna come valore della soglia superiore 209 giorni equivalenti a circa 7 mesi. Applicando tale valore di soglia all'intera banca dati il numero di rapporti la cui data di fine viene modificata è pari all'8% di tutti i rapporti di lavoro interinale chiusi (esclusi i giornalieri). Si è deciso di applicare la chiusura solo ai contratti di lavoro interinale a tempo determinato, non applicando invece alcuna regola ai contratti a tempo indeterminato.

- Lavoro a progetto: il valore della soglia superiore è di 816,5 giorni equivalenti a circa 2 anni. I rapporti presenti in banca dati con durata superiore a tale soglia a cui viene dunque modificata la data di chiusura costituiscono il 6% di tutti i rapporti di lavoro a progetto considerati (esclusi i giornalieri).

- Gestione dei contratti di apprendistato: la gestione dei contratti di apprendistato ha subito modifiche normative nel corso della gestione dei dati ed il processo di trattamento è stato adattato di conseguenza. Fino al novembre 2011 tali contratti sono stati gestiti come normali contratti a termine, prevedendo dunque una data di cessazione allo scadere della quale viene generata la conclusione del rapporto secondo le modlaità descritte in precedenza. A partire da tale data invece il contratto di apprendistato è stato modificato assumendo che al suo termine scatti automaticamente l'assunzione a tempo indeterminato. Le procedure di trattamento quindi gestiscono normalmente i casi di apprendistato fino al novembre 2011, mentre i contratti successivi prevedono non più la generazione di una data di cessazione prevista ma di una data di trasformazione prevista che ipotizza il passaggio al contratto a tempo indeterminato. I primi effetti di tale modifica dovrebbero avere impatto sulla banca dati a partire dal mese di novembre 2013.
- Gestione dei rapporti part-time: il processo di trattamento assume che ad ogni soggetto possa essere associato un unico rapporto lavorativo nel caso in cui si tratti di lavoro a tempo pieno o di contratto con modalità lavorativa non definita. Nel caso in cui il soggetto effettui lavori part time, viene data la possibilità di mantenere aperti contemporaneamente più rapporti part time con imprese differenti. Il sistema è parametrico e può quindi gestire un numero arbitrario K di rapporti part-time contemporanei. Attualmente il processo ad oggi in esercizio, sulla base delle osservazioni effettuate, fissa il valore di K = 2.



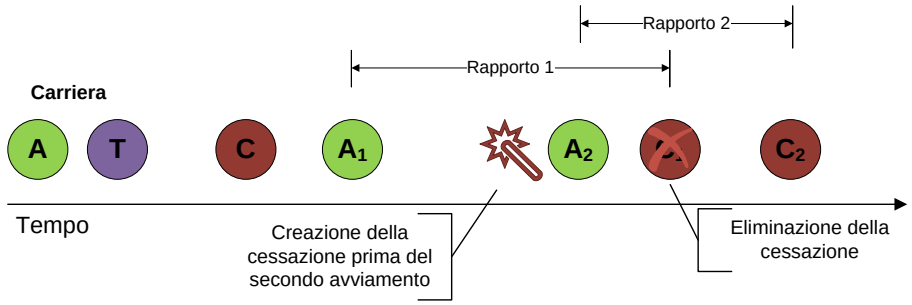
La modalità di correzione introdotta all'interno del processo di messa in qualità del dato impone la chiusura di tutti i rapporti part time ogni volta che il soggetto in esame inizia un contratto di lavoro full time.



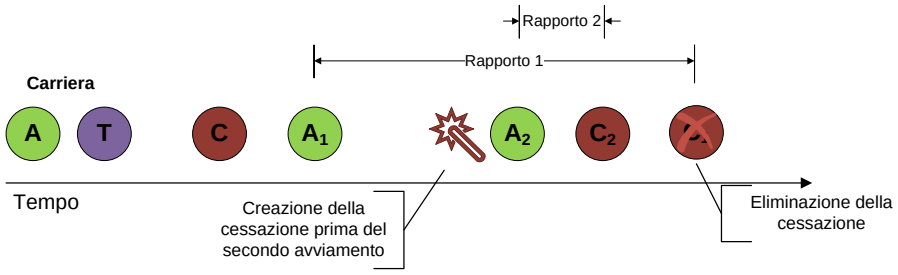
Per garantire questo controllo di coerenza sulle carriera è assolutamente necessario che il campo modalità di lavoro delle comunicazioni obbligatorie sia valorizzato,

nel caso in cui questo campo non sia valorizzato viene assegnato di default il valore di contratto a tempo pieno.

- Applicazione dei controlli di coerenza: le regole descritte fino a questo punto possono combinarsi tra loro nel corso del trattamento dei rapporti di lavoro, generando a loro volta regole di carattere più generale che possono essere utilizzate sia come verifica dell'applicazione delle regole precedenti, sia come ulteriore trattamento per la gestione di casi che esulano da quelli descritti. Di seguito vengono presentate due regole fondamentali per la gestione dei rapporti nel loro complesso. Per poter essere trattati in maniera omogenea gli eventi vengono innanzitutto ordinati in base alla data di riferimento.
  - Controlli di coerenza fra eventi con sovrapposizione parziale: nel caso di sovrapposizione parziale (in cui cioè la coda del primo evento si sovrapponga alla testa del successivo) l'evento più recente viene considerato maggiormente affidabile e il termine dell'evento precedente viene spostato al giorno precedente all'inizio del successivo. Come di consueto il posizionamento della cessazione introdotta può essere configurato in base alle esigenze. Esempi di eventi con sovrapposizione temporale ammessa sono due eventi a tempo parziale di 4 ore al giorno ciascuno. Esempi di eventi con sovrapposizione temporale non ammessa sono due eventi a tempo pieno.



  - Controlli di coerenza fra eventi con sovrapposizione totale: nel caso di sovrapposizione totale (in cui cioè un evento sia interamente contenuto in un altro) l'evento più recente viene considerato maggiormente affidabile e il termine dell'evento precedente viene spostato al giorno precedente all'inizio del successivo.



- Recupero delle qualifiche: il processo di messa in qualità verifica, oltre che la coerenza temporale fra gli eventi che costituiscono una carriera, anche la coerenza tra le informazioni riportate al loro interno. La qualifica di lavoro è uno dei campi che talvolta non è valorizzato all'interno di una comunicazione. Per ovviare a tale

problema nel corso del trattamento nell'ambito dello stesso rapporto di lavoro si cerca di propagare la qualifica, se presente in almeno un evento, a tutti gli eventi con qualifica mancante. Se il campo risulta non valorizzato in un evento viene assunta come valida la qualifica dell'evento precedente o del primo evento precedente con qualifica valorizzata. La modalità di imputazione è dunque una propagazione in avanti nel tempo della qualifica fino a che non si riscontra una valorizzazione, uguale o diversa dalla precedente, della qualifica. Si noti che, nella maggior parte dei casi, la propagazione della qualifica avviene utilizzando una qualifica già presente nello stesso rapporto di lavoro.

- Controlli di consistenza effettuati sulle tipologie contrattuali e sui titoli di studio: sono stati implementati alcuni controlli di consistenza legati alle tipologie contrattuali e ai titoli di studio in relazione all'età dei lavoratori. Le procedure introdotte verificano che le seguenti tipologie contrattuali
  - Apprendistato
  - Tirocinio
  - Contratto d'inserimento

  e che i seguenti titoli di studio

  - Titolo di accademia di belle arti
  - Conseguimento del dottorato di ricerca
  - Conseguimento del diploma universitario
  - Conseguimento della laurea (vecchio e nuovo ordinamento)
  - Formazione post universitaria
  - Istituto professionale e scuola superiore

  siano compatibili con l'età anagrafica del lavoratore analizzato. In caso di incompatibilità il contratto o il titolo vengono posti a valore nullo.

- Verifica della modalità di lavoro riportata nelle cessazioni: poiché la cessazione è una comunicazione che non prevede una variazione della modalità di lavoro di un contratto, è stato inserito un controllo per verificare la veridicità di tale assunto. Nel caso in cui la cessazione riporti al suo interno una modalità lavorativa differente dalla comunicazione che la precede, il valore riportato dalla cessazione viene considerato errato e aggiornato con il valore della comunicazione precedente.
- Gestione unificata dell'anagrafica dei lavoratori: le comunicazioni non prevedono riferimenti ad anagrafiche univoche e riportano interamente tutte le informazioni dei soggetti interessati. Se da un lato ciò consente la massima flessibilità nel corso della comunicazione, dall'altro non permette di verificare la consistenza delle informazioni riferite allo stesso soggetto tra le diverse comunicazioni. Per ovviare a tale problema nel corso del processo viene generata una anagrafe storicizzata dei lavoratori interessati da comunicazioni: la storicizzazione consente di tener traccia delle modifiche ai campi che ammettono una variazione temporale (comune di domicilio, ecc.) mentre la gestione univoca dell'anagrafica consente di verificare la persistenza dei campi immutabili (genere, data di nascita, ecc.). L'introduzione di versioni storiche comporta la presenza in anagrafica di più record riguardanti lo stesso soggetto, caratterizzati dallo stesso identificativo, ma da diversi numeri (sequenziali) di versione. Ciascun soggetto viene dunque riferito all'interno di una

comunicazione non più solo col proprio identificativo ma anche col numero progressivo correlato.

- Recupero del genere e della data di nascita: dopo aver verificato la presenza in banca dati di soggetti con differenti date di nascita o differenti generi, ed essendo queste due caratteristiche invarianti nel tempo, si è deciso di applicare un controllo di coerenza su questi due campi al fine di rendere uniforme l'informazione. In particolare per ogni soggetto presente in banca dati viene controllato quanti valori distinti, diversi dal dato mancante, sono contenuti nei due campi sopracitati e le occorrenze. Una volta recuperate queste due informazioni il valore più frequente viene propagato sull'intera banca dati.

- Gestione unificata dell'anagrafica delle aziende: analogamente a quanto avviene per i lavoratori anche le informazioni delle aziende concorrono alla generazione di una anagrafe storicizzata che registri i cambiamenti riferiti alla stessa unità operativa nel tempo.
- Integrazione dei dati territoriali: il processo di trattamento ed integrazione dei dati, oltre che garantire le correttezza e la coerenza delle informazioni, ha anche l'obiettivo di gestire l'integrazione territoriale dei dati. Come descritto in precedenza infatti le fonti informative coinvolte nel processo sono in parte a carattere regionale, in parte a carattere provinciale; inoltre la competenza del dato riportato può essere ricondotta al livello provinciale sia per mezzo del domicilio del lavoratore sia per mezzo della sede aziendale oggetto di comunicazione. In passato la gestione del dato all'interno di sistemi diversi ha causato disallineamenti tra le diverse sorgenti con la conseguente difficoltà nella produzione di analisi omogenee. L'adozione di un sistema centralizzato dal punto di vista del trattamento, ma federato dal punto di vista della distribuzione, consente di ovviare a questo problema: il trattamento a livello integrato (regionale) consente di eliminare eventuali istanze duplicate e di gestire correttamente la successiva fase di reinstradamento degli eventi alle province di competenza. Al termine del processo di trattamento vengono dunque creati diversi datamart provinciali che riportano i contenuti di competenza di ciascuna provincia: dal punto di vista algebrico la numerosità dei record integrati è maggiore della somma dei singoli datamart (poiché un evento può essere di competenza di province diverse, per lavoratore o per azienda), dal punto di vista logico il datamart regionale rappresenta l'unione dei singoli datamart provinciali e le numerosità tra i due livelli coincidono.

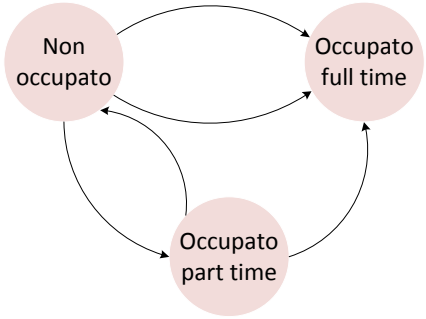
4.2 Il modello formale

In estrema sintesi i trattamenti descritti nel precedente paragrafo e finalizzati alla ricostruzione di carriere lavorative coerenti si riconducono ad una modellazione della carriera lavorativa di un lavoratore come un processo a stati finiti in cui il soggetto può occupare tre stati:

- Occupato full time
- Occupato part time
- Non occupato

In funzione della presenza in uno di questi stati esistono una serie di eventi ammissibili che portano alla transizione in un altro stato e una serie di eventi non ammissibili. Ad

esempio se un lavoratore è nello stato “non occupato” è ammissibile un avviamento che lo porta nello stato “occupato” (full o part time). Non è invece ammissibile una cessazione: per poterla ammettere è necessario generare un avviamento precedente o eliminare la cessazione dall’elaborazione.



Le regole di business descritte in precedenza hanno lo scopo di regolare la transizioni tra i diversi stati e, in caso di transizione non ammissibile, di effettuare le opportune modifiche per fare in modo che la transzione sia ammissibile.

Questa modellizzazione alla base del processo di messa in qualità verrà descritta nel dettaglio e in modo formale nel seguito del documento. La sua introduzione ha lo scopo di favorire la comprensione delle modalità di applicazione delle regole descritte in precedenza e degli esempi presentati nel prossimo paragrafo.

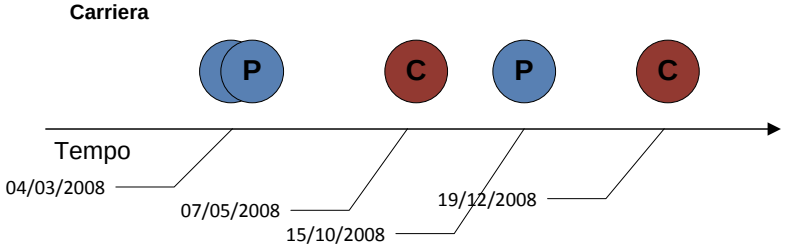
4.3 Alcuni esempi di trattamento

Per poter meglio comprendere le modalità con cui vengono utilizzate le regole descritte in precedenza è possibile analizzare come alcune carriere vengono messe in qualità e qual è il risultato del trattamento. Gli esempi presentati sono estratti direttamente dai casi realmente trattati.

Prendiamo come esempio la successione di eventi presentata in figura:

Data inizio	Data fine	Data cessazione	Rapporto	Tipo evento	Azienda
	04/03/2008	05/04/2008	Tempo determinato	Proroga	Azienda 1
	04/03/2008	05/04/2008	Tempo determinato	Proroga	Azienda 1
05/04/2008	07/05/2008		Tempo determinato	Cessazione	Azienda 1
	15/10/2008	09/12/2008	Tempo determinato	Proroga	Azienda 1
09/12/2008	19/12/2008		Tempo determinato	Cessazione	Azienda 1

È evidente che sono presenti al suo interno eventi non coerenti fra di loro: il rapporto viene prorogato, ma mai avviato.

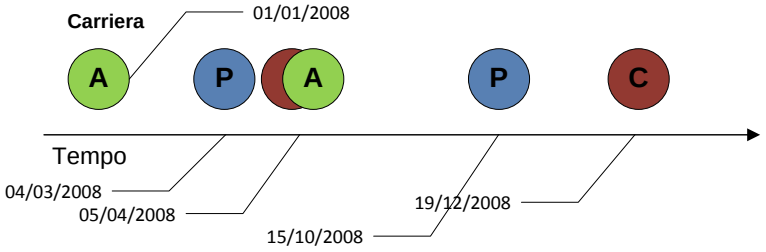


Nella tabella sottostante viene riportato il risultato al termine del processo di trattamento.

Data evento	Rapporto	Tipo evento	Azienda
01/01/2008	Tempo determinato	Avviamento	Azienda 1
04/03/2008	Tempo determinato	Proroga	Azienda 1
05/04/2008	Tempo determinato	Cessazione	Azienda 1
05/04/2008	Tempo determinato	Avviamento	Azienda 1
15/10/2008	Tempo determinato	Proroga	Azienda 1
19/12/2008	Tempo determinato	Cessazione	Azienda 1

La carriera è ora formalmente corretta dopo l'applicazione delle seguenti modifiche:

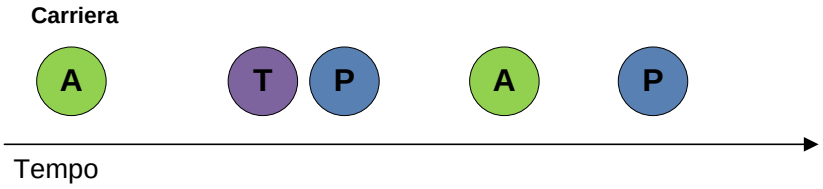
- è stata eliminata la doppia proroga;
- dagli eventi iniziali sono state estratte le informazioni riferite ad altri eventi (ad esempio la cessazione nella prima proroga o il secondo avviamento dalla cessazione);
- sono stati generati gli eventi mancanti posizionandoli nel momento statisticamente più probabile (ad esempio il primo avviamento).



Si consideri questo secondo esempio:

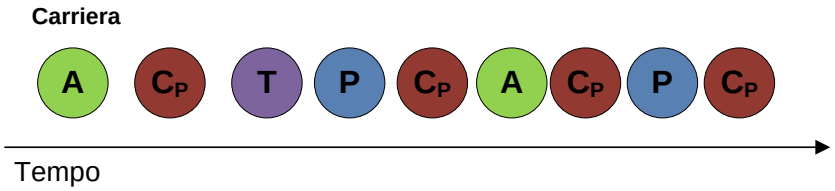
Data inizio	Data fine	Inizio proroga	Fine proroga	Data trasformazione	Rapporto	Tipo evento	Azienda
21/11/2008	30/11/2008	24/11/2008			Tempo determinato	Avviamento	Azienda 1
		04/12/2008		04/12/2008	Tempo determinato	Trasformazione	Azienda 1
		16/12/2008	15/01/2009		Tempo determinato	Proroga	Azienda 1
16/01/2009	28/02/2009	20/01/2009			Tempo determinato	Avviamento	Azienda 1
		17/02/2009	31/05/2009		Tempo determinato	Proroga	Azienda 1

Anche in questo caso è evidente che la carriera non è consistente: mancano alcune cessazioni (se trattasi di tempo determinato full-time) che chiudano i rapporti.

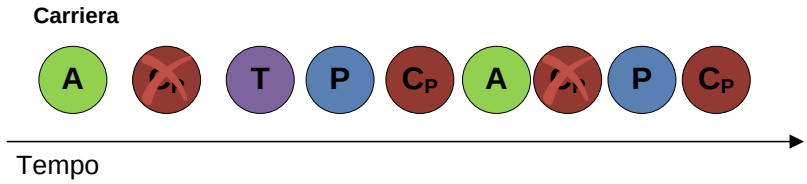


Il primo passaggio di estrazione degli eventi puntuali produce il seguente risultato:

Data evento	Rapporto	Tipo evento	Sottotipoevento	Azienda
21/11/2008	Tempo determinato	Avviamento		Azienda 1
30/11/2008	Tempo determinato	Cessazione	Cessazione prevista	Azienda 1
04/12/2008	Tempo determinato	Trasformazione		Azienda 1
16/12/2008	Tempo determinato	Proroga		Azienda 1
15/01/2009	Tempo determinato	Cessazione	Cessazione prevista	Azienda 1
16/01/2009	Tempo determinato	Avviamento		Azienda 1
28/02/2009	Tempo determinato	Cessazione	Cessazione prevista	Azienda 1
17/02/2009	Tempo determinato	Proroga		Azienda 1
31/05/2009	Tempo determinato	Cessazione	Cessazione prevista	Azienda 1

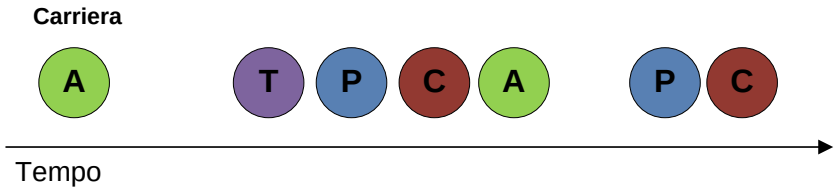


Gli eventi sembrano ora più completi (sono state create le cessazioni previste a partire dagli avviamenti ve dalle proroghe) ma è evidente come la prima cessazione non tenga conto della successiva proroga.



Il secondo passaggio di messa in qualità per mezzo dell'applicazione delle regole di business produce infine il seguente risultato:

Data evento	Tipo rapporto	Tipo evento	Azienda
21/11/2008	Tempo determinato	Avviamento	Azienda 1
04/12/2008	Tempo determinato	Trasformazione	Azienda 1
16/12/2008	Tempo determinato	Proroga	Azienda 1
15/01/2009	Tempo determinato	Cessazione	Azienda 1
16/01/2009	Tempo determinato	Avviamento	Azienda 1
17/02/2009	Tempo determinato	Proroga	Azienda 1
31/05/2009	Tempo determinato	Cessazione	Azienda 1

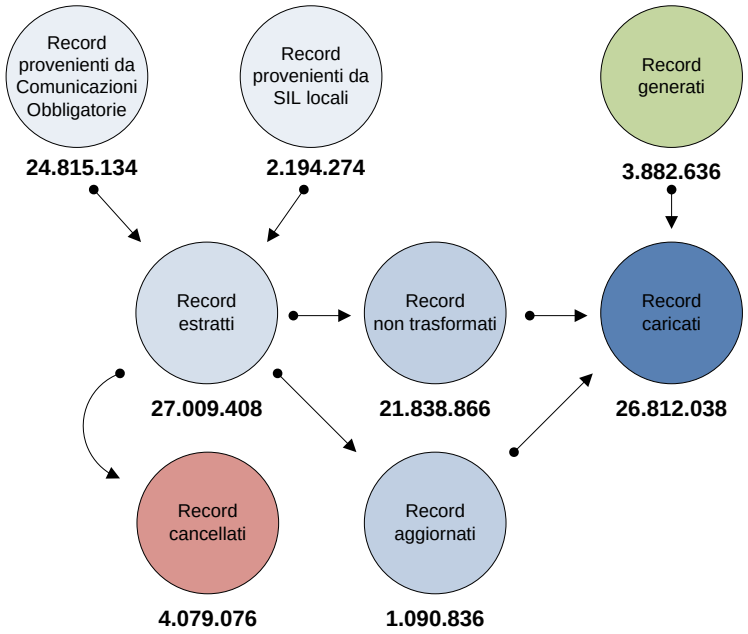


Sono state eliminate le due cessazioni previste in seguito annullate dalla proroga e la carriera risulta ora formalmente corretta.

4.4 L'impatto del trattamento

Per agevolare la comprensione dei processi descritti e valutare il loro impatto sui dati nel loro complesso vengono elencate nel seguito alcune statistiche descrittive riguardanti i record coinvolti nelle diverse fasi del processo. I dati riportati si riferiscono al mese di Settembre 2012.

I record provenienti da SIL locali riguardano eventi successivi al Gennaio 2007 o, se precedenti, relativi solo ad individui che hanno avuto movimentazione dopo il Gennaio 2007.



In figura vengono riportati i macro flussi numerici del processo di trattamento che consentono di evidenziare i record caricati, i record sottoposti a trattamento, i record generati e i record inseriti. Si noti come anche se le numerosità iniziale e finale sono molto simili (circa 27 milioni di record) ciò è l'effetto di due flussi uguali ed opposti che portano alla cancellazione ed alla generazione di un volume simile di record, corrispondente a circa al 15% della banca dati finale, a cui si va ad aggiungere il 5% di comunicazioni aggiornate.

La tabella sottostante riporta invece la numerosità dei record interessati dai principali trattamenti effettuati. La statistica viene effettuata sul data finale e quindi comprende i soli trattamenti che comportano la generazione di record o il loro aggiornamento.

Trattamento	Prima del 2008	2008	2009	2010	2011	2012	Dopo il 2012	Totale
Aggiornamento della modalità di lavoro	120.463	101.355	152.103	191.189	276.135	225.067	24.524	1.090.836
Avviamento creato da cessazione	22.198	65.656	22.404	21.074	15.423	13.032	5.835	165.622
Avviamento creato da proroga o da trasformazione	97.429	137.683	56.117	63.029	63.022	49.604	3.011	469.895
Avviamento iniziale creato perché mancante	331.466	0	0	0	0	0	0	331.466
Cessazione creata perché mancante	722.951	430.951	255.928	275.856	297.224	201.521	4.562	2.188.993
Cessazione create scadenza durata massima	79.982	44.318	59.224	76.929	97.151	113.640	255.116	726.360
Nessun trattamento	5.240.363	3.590.118	3.285.348	3.410.494	3.619.274	2.555.630	137.639	21.838.866
Totale	6.614.852	4.370.081	3.831.124	4.038.571	4.368.229	3.158.494	430.687	26.812.038

5. Validazione formale della consistenza del processo di ETL

Come descritto nelle sezioni precedenti, la procedura di ETL, nella fase di trasformazione del dato, permette di incrementare la **consistenza,** la **correttezza** e la **completezza** dei dati in un database sorgente (S) attraverso l'implementazione di business rule, permettendo quindi la generazione di un *nuovo* dataset (N) corretto, completo e consistente.

In questa sezione di descrive un approccio per la verifica della **consistenza** dei dati, la **Robust Data Quality Analysis** (RDQA), che ha lo scopo di validare il grado di consistenza raggiunto dalle business rule realizzate all'interno del processo di ETL.

Per la realizzazione di una delle componenti della RDQA si è deciso di utilizzare un approccio basato sui metodi formali (ed in particolare il Model Checking). Più precisamente, il Model Checking (tecnica formale per la verifica di proprietà invarianti su sistemi complessi) ha permesso di:

1. Utilizzare gli Automi a Stati Finiti per la modellazione del dominio delle Comunicazioni Obbligatorie e regole di consistenza sui dati;
2. Utilizzare gli algoritmi di model checking per la verifica della consistenza su database di grandi dimensioni;
3. Analizzare la consistenza dei dati ottenuta dai due approcci (approccio formale e basato su business rule), per garantire una maggior efficacia delle operazioni di messa in qualità, individuando e correggendo gli errori.

Di seguito sarà descritta la RDQA istanziata per il dominio del mercato del lavoro descritto nel Capitolo 3. Chiaramente la tecnica è *domain-independent* e può essere quindi applicata ad un generico dominio di data quality.

Si noti, inoltre, che per chiarezza espositiva alcuni dettagli formali (es., definizioni delle funzioni e degli algoritmi della RDQA) sono state omesse. A tal proposito, si rimanda all'articolo (Mezzanzanica, Boselli, Cesarini, & Mercorio, 2011). Infine, il modello di consistenza qui descritto è stato ulteriormente sviluppato per svolgere un'analisi di sensitività sugli effetti delle procedure di messa in qualità applicate ai dati. Per approfondimenti si veda (Mezzanzanica, Boselli, Cesarini, & Mercorio (a), 2012) e (Mezzanzanica, Boselli, Cesarini, & Mercorio (b), 2012)

Per rendere più chiara la metodologia utilizzata, di seguito si descriverà informalmente il concetto di Automa a Stati Finiti e di Model Checking su Automi a Stati Finiti. Successivamente si descriverà la tecnica della Robust Data Quality Analysis.

5.1 Automi a Stati Finiti

Un Automa a Stati Finiti (di seguito FSS) è un modello basato principalmente sul concetto di ***stato*** e ***transizione tra stati*** ed ha lo scopo di modellare la dinamica di un sistema che evolve nel tempo in funzione degli eventi che questo riceve come input.

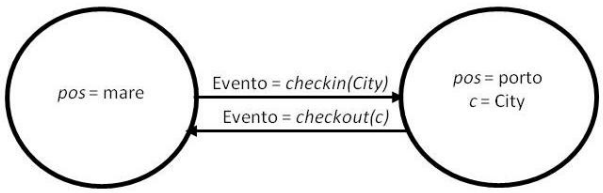
Lo stato descrive, in termini di **variabili** e **valori**, quali sono gli elementi che compongono il sistema, mentre una transizione tra stati rappresenta il legame esistente tra due stati (anche non distinti) mediante una data azione. Genericamente, una transizione tra due stati indica che il sistema può transire da uno stato s_i ad uno stato s_j tramite un'azione a_k se le pre-condizioni per l'esecuzione di a_k sono verificate nello stato s_i.

Si pensi all'esempio della navigazione di Tabella 1. Un evento generico e_i, in questo caso, è identificato da una selezione sugli attributi $e_i = (ShipID_i, Città_i, Data_i, TipoEvento_i)$. Infine, è possibile ordinare gli eventi in base alla data (ossia, $\forall e_i, e_j \in E, e_i \leq e_j$ iff $Data_{e_i} \leq Data_{e_j}$). La frase *"tutte le volte che una nave fa un checkin in un porto farà un checkout prima di entrare in un nuovo porto"* è un esempio di proprietà di consistenza che può essere verificata usando gli FSS. L'Automa a Stati Finiti di questo esempio può essere modellato come descritto nella Figura successiva.

- Lo "stato" del sistema, in questo caso, è definito da due variabili distinte:
- Variabile *pos*, che definisce la posizione della nave e che può assumere rispettivamente valori *pos=mare* quando la nave è in mare, *pos=porto* quando la nave ha attraccato al porto.
- Variabile *c= nome della città* che modella la città nella quale si trova la nave.

Le *transizioni*, cioè gli eventi che alterano lo stato del sistema, sono gli eventi di *checkin(Città)* e *checkout(Città)* che rispettivamente fanno transire il sistema da uno stato all'altro. La Figura mostra una possibile modellazione dell'FSS in cui gli stati descrivono l'evoluzione consistente del dominio descritto.

Ad esempio, una sequenza di eventi come quella descritta in Tabella (ossia, checkin(Venezia), checkout(Venezia), checkin(Genova), checkin(Genova)) è consistente solo per i primi tre eventi, mentre il quarto evento (in rosso) rende *tutta* la sequenza inconsistente, poiché questo non può essere applicato a nessuna transizione dell'automa.

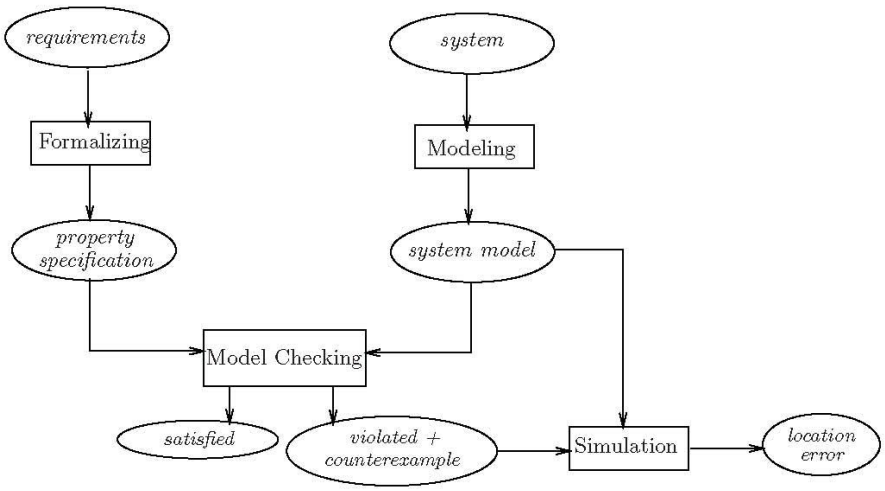


5.2 Breve Descrizione del Model Checking su FSS

Il Model Checking (si veda (Clarke, Grumberg, & Peled, 1999) (Baier & Katoen, 2008) per approfondimenti) è una tecnica automatica per la verifica di modelli di sistemi hardware/software. In particolare, il Model Checking prevede la definizione di un *modello* che descriva l'evoluzione del sistema in base al tempo (**System Modelling**). Questo modello, in genere, è descritto tramite FSS (nel caso del Model Checking *esplicito*) come descritto nella sezione precedente. Successivamente si formalizza la proprietà che si vuole verificare (espressa usando o la logica proposizionale o la logica temporale).

Un model checker, cioè il software che implementa la tecnica del Model Checking, verifica esaustivamente che tutte le possibili configurazioni in cui il sistema può giungere, a partire da una condizione iniziale specificata, soddisfino sempre la proprietà **(System**

Verification). Facendo ancora un parallelo con l'esempio precedente, il model checker verifica che tutte le sequenze generate dai possibili eventi in input non conducano mai in uno stato inconsistente. In altri termini, un model checker verifica che in tutti i possibili casi di esecuzione del sistema la proprietà sia sempre soddisfatta. In caso affermativo il sistema si può dire corretto (rispetto alla proprietà verificata). Altrimenti, il model checker fornirà un contro esempio, cioè una traccia d'errore che mostra *come* il sistema ha raggiunto una configurazione che falsifica la proprietà. Una descrizione grafica dell'approccio è fornita in Figura (presa da (Baier & Katoen, 2008)).



5.3 Robust Data Quality Analysis

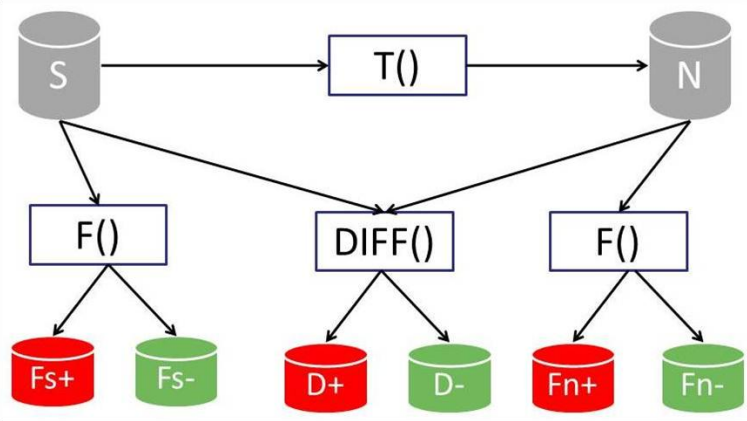
La *Robust Data Quality Analysis* (RDQA) è una tecnica di analisi e misurazione del grado di consistenza di un DB basata sull'utilizzo di due funzioni:

1. la funzione ***T():*** implementa le business rule all'interno del processo di ETL;
2. la funzione ***F():*** frutta il Model Checking per verificare la consistenza del dato prima e dopo l'intervento di ***T().***

In figura è rappresentata schematicamente l'idea della tecnica realizzata. Per chiarezza espositiva, di seguito si farà riferimento esplicito al concetto di Carriera che sarà definito nella Sezione 5.4. Informalmente, una carreria, così come introdotto nei capitoli precedenti, descrive una sequenza temporalmente ordinata di Comunicazioni Obbligatorie riguardanti il medesimo soggetto.

1. Il dato sorgente ed inconsistente *S* viene fornito in input al processo di messa in qualità T(S). Il prodotto risultante è un dataset *N* in cui le inconsistenze riscontrate sono trattate secondo quanto descritto nella documentazione del processo di ETL.
2. Successivamente, il dato sorgente *S* viene validato usando l'approccio basato su model checking F(S). Il risultato non sarà un nuovo dataset corretto, ma bensì una *partizione* del dataset sorgente in carriere inconsistenti su *S* (indicate con Fs+) e consistenti su *S* (indicate con Fs-).
3. Il procedimento al punto 2 viene iterato usando come parametro di ingresso il dataset *N*, ossia il database ottenuto come risultato del processo di T(S). Si ottiene quindi il partizionamento del dataset successivo al trattamento ETL rispettivamente in carriere inconsistenti (Fn+) e consistenti (Fn-).

4. Col fine di ottenere un più efficace paragone tra i dataset pre/post ETL, la procedura DIFF() suddivide l'insieme delle carriere tra quelle che hanno subito almeno un alterazione durante il processo ETL (D+) e quelle che non hanno subito alcuna alterazione (D-).
5. Infine, valutando la combinazione dei valori di Fs+,Fs-,D+,D-,Fn+,Fn- è possibile analizzare come ogni singola carriera è stata trattata dal processo di ETL. Il risultato è una **Double Check Matrix** (DCM) che classifica una carriera X in uno degli otto possibili gruppi, come raffigurato in tabella.



	A quale di queste partizioni appartiene la carriera X?						Descrizione
Gruppo	***Fs+***	***Fs-***	***D+***	***D-***	***Fn+***	***Fn-***	
1	NO	SI	NO	SI	NO	SI	Carriera consistente e non alterata da T().
2	NO	SI	NO	SI	SI	NO	Carriera inizialmente consistente, non alterata da T() ma erroneamente considerata consistente da T(), potrebbe indicare un errore nell'implementazione della F() o della DIFF().
3	NO	SI	SI	NO	NO	SI	Carriera consistente sia prima che dopo, ma alterata nella sostanza da T(). Potrebbe generare una inconsistenza futura.
4	NO	SI	SI	NO	SI	NO	Carriera inizialmente consistente ma alterata da T() e resa inconsistente.
5	SI	NO	NO	SI	NO	SI	Carriera inizialmente inconsistente, non alterata da T() e diventata consistente. Se presente è indice di un bug in T().
6	SI	NO	NO	SI	SI	NO	Carriera inizialmente inconsistente che T() non è stata in grado di individuare e correggere.
7	SI	NO	SI	NO	NO	SI	Carriera inizialmente inconsistente che T() ha corretto efficacemente.
8	SI	NO	SI	NO	SI	NO	Carriera inizialmente inconsistente sulla quale T() è intervenuta, tuttavia l'intervento ha prodotto un nuovo errore che continua a rendere la carriera inconsistente.

La **Double Check Matrix (DCM)** ha lo scopo di validare le regole di consistenza di T() sulla base dell'analisi formale di F() e viceversa, fornendo uno strumento iterativo,

strutturato e tracciabile, utile al miglioramento dell'intero processo di messa in qualità del dato.

Di seguito si illustrerà la semantica modellata dalla funzione formale F() nel dominio di interesse.

5.4 Modellazione della Funzione Formale F() per il Mercato del Lavoro

Nel contesto dell'analisi del mercato del lavoro, un evento (una Comunicazione Obbligatoria) è composto principalmente dalle seguenti informazioni:

- **worker_id:** è l'identificato della persona coinvolta nell'evento;
- **e_id:** è l'identificato dell'evento;
- **e_date:** rappresenta la data in cui l'evento avviene;
- **e_type:** nelle carriere lavorative il tipo evento può assumere uno tra i seguenti valori:
  - *Avviamento (st)*: l'individuo ha avviato un nuovo contratto di lavoro;
  - *Proroga (ex)*: l'individuo ha effettuato una proroga del contratto di lavoro già attivo;
  - *Trasformazione (cn)*: l'individuo ha trasformato il suo contratto di lavoro in un altro con modalità o tipologia di rapporto differente;
  - *Cessazione (cs)*: l'individuo ha effettuato una cessazione del suo contratto di lavoro.
- **c_flag:** identifica se il contratto è attivo in modalità part-time (PT) o full-time (FT);
- **c_type:** individua le caratteristiche del contratto in accordo con la legge italiana (es., contratto a tempo determinato, contratto indeterminato, contratto di apprendistato, etc…).
- **empr_id:** identifica univocamente l'azienda alla quale il contratto si riferisce.

Più precisamente, nel contesto dell'analisi del mercato del lavoro possiamo definire:

- **Evento**: una tupla $e_i =$ (worker_id, e_id, e_date, e_type, c_flag, c_type, empr_id).
- **Carriera**: sequenza finita ed ordinata di eventi $e_1, e_2, \dots, e_n$ che descrive lo stato lavorativo dell'individuo all'istante della ricezione dell'evento *n.*

L'evoluzione della carriera rispetto al tempo è sempre ordinata in funzione della data di arrivo dell'evento. Un evento modella ogni singola modifica/variazione dello stato lavorativo degli individui pervenuta attraverso una comunicazione obbligatoria. Tutti gli eventi associati al medesimo individuo ne compongono la ***carriera lavorativa***.

Alcune delle proprietà che descrivono l'evoluzione consistente di una generica carriera sono:

C1. Un dipendente non può avere attivo più di un contratto a tempo pieno;

C2. Un dipendente può avere attivi al massimo *k* contratti part-time (firmati con aziende diverse)[*];

C3. Un lavoratore non può avere attivi contestualmente un contratto part-time e un full-time;

C4. Non si può effettuare una proroga di un contratto a tempo indeterminato;

C5. Una proroga di contratto si può effettuare solo se la tipologia (c_type) e la modalità di contratto (c_flag) non cambiano;

C6. Una trasformazione di contratto richiede che o la tipologia (c_type) o la modalità di contratto (c_flag) cambiano (o entrambe).

La carriera di un lavoratore ad un determinato momento storico (ossia, lo stato) nel nostro modello di validazione della consistenza impiegato può essere identificata dalle seguenti variabili:

VARIABILE 1. C[] è la lista delle aziende con le quali il lavoratore ha un contratto attivo;

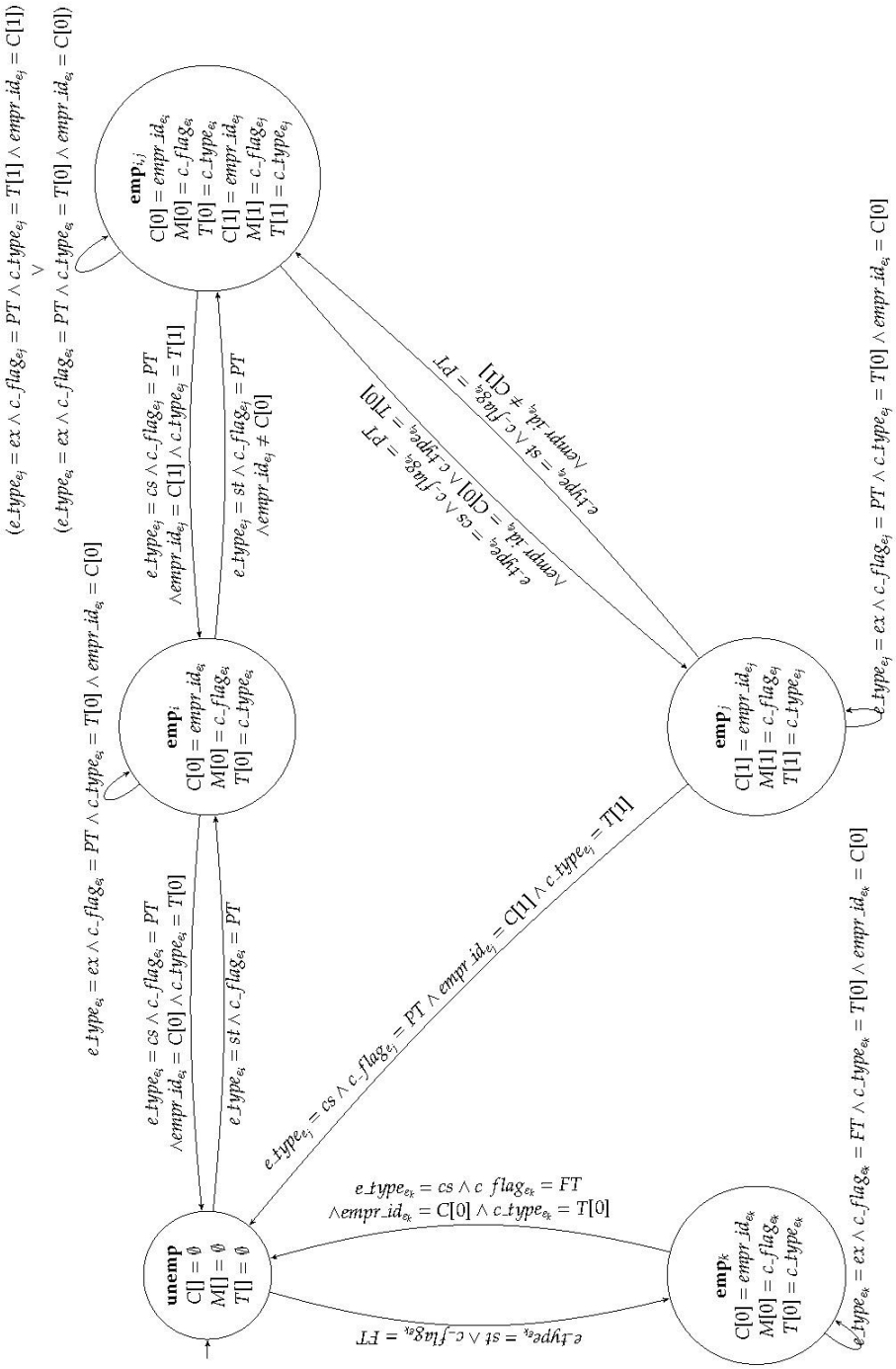
VARIABILE 2. M[] è la lista delle modalità (c_type) con la quale il lavoratore ha attivo il contratto per ognuno dei rapporti di lavoro attivi;

VARIABILE 3. T[] è la lista delle caratteristiche del contratto (c_type) che descrivono il contratto attivo per ognuno dei contratti attivi.

A titolo esemplificativo, uno stato in un dato momento temporale composto da C[0] = 12, M[0] = PT e T[0] = *"indeterminato"* indica che il lavoratore ha attivo un contratto part-time a tempo indeterminato con l'azienda avente id 12.

Per completezza, in Figura si mostra il FSS che modella il dominio appena introdotto, utilizzato dal Model Checker CMurphi per la verifica delle carriere.

[*]Il valore *k* è parametrico e può essere istanziato a piacimento. In questo lavoro, per esigenze di analisi, si è scelto di fissare k=2.


unemp
C[] = ∅
M[] = ∅
T[] = ∅
emp_i
C[0] = empr_id_ei
M[0] = c_flag_ei
T[0] = c_type_ei
emp_j
C[1] = empr_id_ej
M[1] = c_flag_ej
T[1] = c_type_ej
emp_k
C[0] = empr_id_ek
M[0] = c_flag_ek
T[0] = c_type_ek
emp_i,j
C[0] = empr_id_ei
M[0] = c_flag_ei
T[0] = c_type_ei
C[1] = empr_id_ej
M[1] = c_flag_ej
T[1] = c_type_ej
e_type_ei = cs ∧ c_flag_ei = PT ∧ empr_id_ei = C[0] ∧ c_type_ei = T[0]
e_type_ei = st ∧ c_flag_ei = PT
e_type_ej = cs ∧ c_flag_ej = PT ∧ empr_id_ej = C[1] ∧ c_type_ej = T[1]
e_type_ej = st ∧ c_flag_ej = PT ∧empr_id_ej ≠ C[0]
e_type_ei = cs ∧ c_flag_ei = PT ∧ empr_id_ei = C[1] ∧ c_type_ei = T[1]
e_type_ei = st ∧ c_flag_ei = PT
e_type_ej = cs ∧ c_flag_ej = PT ∧ empr_id_ej = C[0] ∧ c_type_ej = T[0]
e_type_ej = ex ∧ c_flag_ej = PT ∧ c_type_ej = T[0] ∧ empr_id_ei = C[1])
∨
(e_type_ei = ex ∧ c_flag_ei = PT ∧ c_type_ei = T[0] ∧ empr_id_ej = C[0])
e_type_ei = ex ∧ c_flag_ei = PT ∧ c_type_ei = T[0] ∧ empr_id_ei = C[0]
e_type_ej = ex ∧ c_flag_ej = PT ∧ c_type_ej = T[0] ∧ empr_id_ej = C[0]
e_type_ei = cs ∧ c_flag_ei = PT ∧ empr_id_ei = C[1] ∧ c_type_ei = T[1]
e_type_ek = cs ∧ c_flag_ek = FT ∧ c_type_ek = T[0] ∧ empr_id_ek = C[0]
e_type_ek = st ∧ c_flag_ek = FT
e_type_ej = ex ∧ c_flag_ej = PT ∧ c_type_ej = T[0] ∧ empr_id_ej = C[0]
e_type_ek = ex ∧ c_flag_ek = FT ∧ c_type_ek = T[0] ∧ empr_id_ek = C[0]

6. Double Check Matrix su DB Lombardia

La metodologia RDQA è stata applicata ad un database di 47.154.010 di comunicazioni obbligatorie riguardanti 5.570.991 di individui, intercorsi in aziende aventi sede operativa in Regione Lombardia ed osservate dal 1° gennaio 2000 al 31 dicembre 2010.

6.1 Double Check Matrix

La RDQA è stata applicata sul database del mercato del lavoro applicando le funzioni *T()*e *F()* come descritto nel Paragrafo 5.3, ottenendo come risultato la Double Check Matrix.

È importante notare che la consistenza viene valutata in termini di **carriere** nonostante il database sia composto da **eventi**. Tuttavia, per ogni gruppo della DCM è anche mostrata una colonna che fornisce informazioni sul numero di eventi appartenenti al gruppo o, in altri termini, quanti sono gli eventi associati alle carriere ricondotte allo specifico gruppo. Una volta ottenuta tale informazione, è possibile *ponderare* il peso, in termini di eventi, che ogni gruppo ha sul database mediante la variabile *'% Ponderata'*. In altri termini, la variabile mostra la percentuale ponderata del numero medio di eventi relativi alle carriere dei singoli gruppi.

Gruppo	Riga	Consi-stente Pre ETL?	Alte-rata da ETL?	Consi-stente Post ETL?	# Carriere	%	# Eventi	% Ponderata
1	1	SI	NO	SI	1.984.692	35,63	5.367.306	12,40
2	2	SI	NO	NO	0	0	0	0
3	3	SI	SI	SI	334.097	6,00	1.429.170	3,16
--	4	SI	SI	null	40.520	0,73	126.644	0,28
4	5	SI	SI	NO	1.100	0,02	9.530	0,02
5	6	NO	NO	SI	0	0	0	0
6	7	NO	NO	NO	15.267	0,27	86.364	0,20
7	8	NO	SI	SI	2.858.357	51,31	34.399.674	71,85
8	9	NO	SI	NO	284.569	5,11	5.314.062	11,14
--	10	NO	SI	null	52.389	0,93	421.260	0,95
		TOTALI			5.570.991	100	47.154.010	100

Nel caso specifico la DCM contiene delle righe anomale, rispettivamente la riga 4 (SI,SI,null) e la riga 10 (NO,SI,null) che non sono classificate in alcun gruppo. E' emerso infatti che tali carriere non sono di competenza della regione in esame sia in termini di sede operativa dell'azienda sia in termini di sede del domicilio del lavoratore. In altri termini, le Comunicazioni Obbligatorie di questi lavoratori non dovrebbero essere memorizzate nel database della Regione. Per questo motivo la funzione *T()* li ha *correttamente* esclusi dal nuovo database consistente *N*.

6.1.1 DCM: Analisi

Si commentano ora brevemente i risultati ottenuti:

Gruppo 1: rappresenta le carriere già consistenti presenti nel database. Possiamo vedere che per la Regione Lombardia sono circa il 35%.

Gruppo 2: rappresenta carriere considerate consistenti da *F()* ma inconsistenti dopo l'intervento di *T()*, sebbene questa non le abbia alterate. Come atteso, questo insieme è vuoto.

Gruppo 3: rappresenta le carriere trattate da T() ma valide sia prima che dopo il trattamento. Possiamo vedere che le carriere appartenenti a questo gruppo sono circa il 6%. Questo gruppo merita particolare attenzione poiché mostra delle differenze semantiche nel trattamento dell'inconsistenza tra le funzioni *T()* ed *F()*. Si noti, tuttavia, che tale divergenze sono fisiologiche quando si usano paradigmi di data quality così eterogenei tra loro, come nel caso di *T()* e *F()*.

Gruppo 4: rappresenta le carriere "rese inconsistenti", cioè quelle inizialmente consistenti, trattate poi dalla funzione *ETL* e divenute inconsistenti. Come ci si può aspettare, la numerosità di questo gruppo è molto bassa.

Gruppo 5: rappresenta carriere considerate inconsistenti da *F()* ma consistenti dopo l'intervento di *T()*, sebbene questa non le abbia alterate. Specularmente al Gruppo 2, questo insieme è vuoto.

Gruppo 6: rappresenta le carriere segnalate non consistenti dalla funzione *F()* e non alterate dalla *T()*. Lo studio delle caratteristiche delle carriere appartenenti a questo gruppo è molto utile per il miglioramento della funzione *T()*.

Gruppo 7: rappresenta le carriere consistenti, cioè quelle carriere considerate inizialmente inconsistenti dalla *F* e consistenti dopo l'intervento di correzione da T(). Come si può osservare queste rappresentano più del 50% del totale per la Regione Lombardia.

Gruppo 8: rappresenta le carriere considerate inconsistenti da *F()*, sia prima che dopo l'intervento di T(). Queste sono circa il 5% della totalità e, in altri termini, si tratta di carriere la cui inconsistenza non è stata rilevata da T() ma è invece rilevata dalla funzione *F()*.

Grazie ai risultati presentati nella DCM, in particolare dopo uno studio più approfondito sulle carriere appartenenti ai gruppi 1,3,5,7, è stato possibile migliorare l'implementazione sia *T()* sia di *F()*.

6.1.2 DCM: Approfondimento

Per migliorare la qualità finale del database è stato effettuato uno studio approfondito sul Gruppo 3, presente in percentuale significativa.

Gruppo 3 (SI,SI,SI): Sono state analizzate le caratteristiche delle carriere che compongono questo gruppo. Si è scoperto che la procedura di messa in qualità ha effettuato dei cambi di data ed eliminazioni di eventi di cessazione.

Esempio1:

	Data Evento	Tipo Evento	Modalità Lavoro	Rapporto Evento	Impresa
Pre ETL	42042	Avviamento	FT	A.04.00	605759
	42786	Cessazione	FT	A.04.00	605759
Post ETL	42053	Avviamento	FT	A.04.00	605759

Nell'*Esempio 1* viene eliminato un evento presente nel database iniziale riguardante una cessazione, poiché il contratto a cui l'evento si riferisce è a tempo indeterminato con una data di chiusura "prevista" (ossia, inserita in automatica dal sistema) invece che "certificata" (ossia, inserita a partire da una Comunicazione Obbligatoria). Questa data secondo la funzione *T()* non dovrebbe esistere in un contratto a tempo indeterminato (si presume, infatti, che un contratto a tempo indeterminate non abbia una data di cessazione prevista). Per tale motivo viene eliminata, introducendo tuttavia una *potenziale* inconsistenza futura che la RDQA rileva.

Esempio2:

	Data Evento	Tipo Evento	Modalità Lavoro	Rapporto Evento	Impresa
Pre ETL	41676	Avviamento	FT	B.01.00	32774
	41963	Cessazione	FT	B.01.00	32774
	42037	Avviamento	FT	B.01.00	32774
	42329	Cessazione	FT	B.01.00	32774
Post ETL	41676	Avviamento	FT	B.01.00	32774
	41963	Cessazione	FT	B.01.00	32774
	42037	Avviamento	FT	B.01.00	32774
	42853	Cessazione	FT	B.01.00	32774

Nell'*Esempio 2* si nota una variazione nella *data di chiusura* dell'ultimo contratto nella stessa carriera prima e dopo l'intervento della funzione T(). Infatti T() interviene modificando la data di chiusura del contratto come descritto nel Capitolo 4.

È stato poi effettuato uno studio per valutare l'incidenza di questi due casi all'interno del gruppo ottenendo i seguenti risultati:

	Approfondimento Riga 3 DCM	
	# Carriere	%
TOTALE	334097	6
Cambiamenti di data	238794	71,47
Eliminazione di chiusura	69213	20,72
Cambiamento data ed eliminazione chiusura	256	0,08
Da approfondire	26346	7,89

Si può vedere che per la Regione questi due casi rappresentano circa il 90% del gruppo. Grazie all'applicazione della RDQA si è riusciti a raffinare le funzioni *F()* e *T()* in modo da individuare e gestire queste due tipologie di trattamento delle inconsistenze, ottenendo un incremento sensibile dell'efficacia dell'intero processo di messa in qualità delle carriere.

6.2 Sintesi dei Risultati su DB COB Lombardia

Grazie ai risultati ottenuti dalla Double Check Matrix (presentata al punto 5.1) è possibile analizzare in maniera approfondita il grado di consistenza dei DB utilizzati, effettuando una valutazione della qualità del dato sorgente e del risultato ottenuto grazie alle operazioni di cleansing realizzate dalla funzione T(). Più precisamente, si è analizzato:

1. Il grado di **consistenza iniziale** del dataset originale prima dell'intervento di T() ;
2. Il grado di **consistenza finale** del database raggiunto da T() sul dato originale (rapporto correzioni effettuate su numero di inconsistenze);
3. Il **margine di miglioramento** di T() descrive quanto ancora può migliorare l'efficacia del processo di cleansing (ad esempio, considerando le regole di intervento che introducono inconsistenze nei dati originalmente consistenti);
4. Il **miglioramento ottenuto** stima l'impatto che la funzione di cleansing T() ha avuto sui dati incosistenti.

Nella tabella seguente vengono raccolti i risultati principali:

	REGIONE Lombardia	
	Carriere (%)	Eventi (%)
Consistenza Iniziale	35,65	12,42
Consistenza Finale	86,94	84,25
Margine di Miglioramento	11,4	14,52
Miglioramento Ottenuto	51,29	71,83

Consistenza Iniziale (Gruppo1+Gruppo4, DCM**)** rappresenta il grado di consistenza del DB prima dell'intervento di messa in qualità. Si può notare come il grado di consistenza del DB relativo alla Regione era inizialmente del 35,65% rispetto alle carriere e 12,42% rispetto agli eventi. E' importante osservare che l'indicazione del 12% del database consistente *non* indica necessariamente che il restante 88% è inconsistente, piuttosto indica

che l'88% degli eventi è relativo a carriere che hanno presentato, anche storicamente, almeno una inconsistenza.

Per interpretare correttamente questo risultato è necessario considerare la natura incrementale (e quindi storica) del database del mercato del lavoro che si sta analizzando. Infatti, ogni volta che una carriera viene etichettata come inconsistente questa rimarrà tale nonostante, nel tempo, ad essa continueranno ad essere associate altre comunicazioni obbligatorie. Questi *nuovi* eventi non potranno alterare la condizione di inconsistenza della carriera, tuttavia contribuiranno all'impatto che tale carriera avrà sulla totalità degli eventi inconsistenti (l'88%).

Diversamente, è possibile affermare con certezza che il Database ha un grado di consistenza iniziale (che ricordiamo essere definito sulle carriere e non sugli eventi) del 35%. Questo dato è sufficiente per affermare che il database necessità di un processo di messa in qualità dei dati, prima di utilizzare il suo contenuto per attività di supporto alle decisioni.

Consistenza Finale (Gruppo1+Gruppo7, DCM**)** rappresenta il grado di consistenza del DB finale, cioè dopo l'applicazione della funzione T(). In questo caso si nota che il grado di consistenza finale del DB grazie all'applicazione di T(), e alle migliorie che la RDQA ha permesso di apportare, arriva alla soglia dell'87% contro il 35% iniziale.

Margine di Miglioramento (Gruppo3+Gruppo4+Gruppo6+Gruppo8, DCM**)** rappresenta, idealmente, quanto ancora si può migliorare nell'implementazione della funzione T() del processo di ETL. Anche questo dato va analizzato alla luce di alcuni fattori, tra i quali va considerato che l'approccio formale, per sua natura, è fortemente dipendente dal modello sottostante e, in genere, assai rigoroso. Come conseguenza anche le inconsistenze più superficiali (es., eventi doppi) sono fonte di inconsistenza.

Miglioramento Ottenuto (Gruppo7-Gruppo4, DCM**)** rappresenta la percentuale di miglioramento apportata al DB tramite l'applicazione del processo di messa in qualità. Per la Regione si è ottenuto un miglioramento del grado di qualità del DB del 51,29% rispetto alle carriere, ovvero la funzione T() è riuscita a rendere consistenti il 51,29% delle carriere inconsistenti, e del 71,83% considerando gli eventi (ossia, l'impatto che queste carriere avevano sulla totalità degli eventi del database).

7. Conclusioni e Prospettive di Sviluppo

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, con riferimento alle banche dati della Regione Lombardia. Si è svolta un'analisi della qualità del dato nelle dimensioni di consistenza, accuratezza e completezza, utilizzando tecniche di ETL per la messa in qualità, dettagliando e motivando i singoli criteri di intervento. Successivamente, si è applicato il Model Checking per la verifica della consistenza della base dati rispetto ad un modello opportunamente formalizzato, confrontando accuratamente i risultati derivanti dai due approcci mediante la Robust Data Quality Analysis.

L'applicazione dell'ETL sul database della Regione Lombardia ha permesso di analizzare la qualità del dato sorgente e generare un nuovo dataset qualitativamente superiore nei termini di accuratezza, consistenza e completezza. La RDQA, infine, ha fornito una validazione formale del procedimento svolto mediante l'ETL. Entrambe le tecniche, seppur ancora in fase di evoluzione, hanno evidenziato la forte dipendenza che connette la qualità del dato 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.

Attualmente, oltre al continuo studio delle metodologie per la messa in qualità e alla sperimentazione su altre fonti informative sia isolate sia integrate, si è interessati all'uso dei metodi formali per l'identificazione e la generazione *automatica* delle business rule.

Parte del lavoro qui descritto è stato oggetto di alcune pubblicazioni scientifiche (Mezzanzanica, Boselli, Cesarini, & Mercorio, 2011, 2012(a), 2012(b), ottenendo riscontri positivi dalla comunità accademica internazionale.

Bibliografia

Arenas, M., Bertossi, L. E., & Chomicki, J. (1999). Consistent Query Answers in Inconsistent Databases. (p. 68-79). ACM Press.

Baier, C., & Katoen, J. P. (2008). *Principles of model checking* (Vol. 26202649). MIT press.

Bankier, M. (1999). Experience with the New Imputation Methodology used in the 1996 Canadian census with extensions for future censuses. *Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing.*

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques.* Springer.

Beeri, C., Fagin, R., & Howard, J. H. (1977). A complete axiomatization for functional and multivalued dependencies in database relations. *Proceedings of the 1977 ACM SIGMOD international conference on Management of data*, (p. 47--61).

Chomicki, J. (1992). History-less checking of dynamic integrity constraints. *Proceedings of the Eighth International Conference on Data Engineering*, (p. 557--564).

Chomicki, J., & Marcinkowski, J. (2005). Minimal-change integrity maintenance using tuple deletions. *Information and Computation, 197*, 90--121.

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model Checking.* The MIT Press.

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health and the Nations Health, 36*, 1412--1416.

Fagin, R. (1977). Multivalued dependencies and a new normal form for relational databases. *ACM Transactions on Database Systems (TODS), 2*, 262--278.

Fan, W. (2008). Dependencies revisited for improving data quality. *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, (p. 159--170).

Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and inputation. *Journal of the American Statistical association, 71*, 17--35.

Fisher, C. W., & Kingma, B. R. (2001, December). Criticality of data quality as exemplified in two disasters. *Inf. Manage., 39*, 109--116.

Galhardas, H., Florescuand, D., Simon, E., & Shasha, D. (2000). An Extensible Framework for Data Cleaning. *Proceedings of ICDE '00* (p. 312--312). IEEE Computer Society.

Mayfield, C., Neville, J., & Prabhakar, S. (2009). A Statistical Method for Integrated Data Cleaning and Imputation. *Purdue University, CSD TR-09-008 Technical Report*.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2011). Data Quality through Model Checking Techniques. *Proceedings of Intelligent Data Analysis (IDA), Lecture Notes in Computer Science vol. 7014* (p. 270-281). Springer.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2012). Data Quality Sensitivity Analysis on Aggregate Indicators. *DATA 2012 - Proceedings of the International Conference on Data Technologies and Applications* (p. 97-108). SciTePress.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2012). Toward the use of Model Checking for Performing Data Consistency Evaluation and Cleansing. *The 17th International Conference on Information Quality (ICIQ) (prossima uscita).*

Muller, H., & Freytag, J.-C. (2003). *Problems, Methods and Challenges in Comprehensive Data Cleansing.* techreport.

Ray, I., & Ray, I. (2001). Detecting Termination of Active Database Rules Using Symbolic Model Checking. *Proceedings of the 5th East European Conference on Advances in Databases and Information Systems* (p. 266--279). Springer-Verlag.

Redman, T. C. (1998, February). The impact of poor data quality on the typical enterprise. *Commun. ACM, 41*(2), 79--82.

Redman, T. C. (2001). *Data quality: the field guide.* Digital Pr.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997, May). Data quality in context. *Commun. ACM, 40*, 103--110.

Vardi, M. Y. (1987). Fundamentals of dependency theory. *Trends in Theoretical Computer Science*, 171--224.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5--33.