INAIL
Quaderni di ricerca

# Infortuni sul lavoro. Un modello di lettura (della numerosità) su "open data" dell'Inail

Giacomo Ciriello, Massimo De Felice,
Rosaria Mosca, Massimiliano Veltroni

numero **1** – maggio 2013

I *Quaderni di ricerca* dell'Inail sono lo strumento a disposizione dei ricercatori e dei ruoli professionali dell'Istituto per rendere pubblici i risultati più rilevanti delle loro attività; accolgono anche contributi di ricercatori esterni, che partecipano ai progetti dell'Inail.
Gli autori hanno la piena responsabilità delle opinioni espresse nei *Quaderni*, che non vanno intese come posizioni ufficiali dell'Inail.

I *Quaderni* sono disponibili online all'indirizzo www.inail.it.

Comitato di redazione: Piero Altarocca, Paolo Bragatto, Massimo De Felice,
Benedetta Persechino, Patrizio Rossi.

INAIL – Piazzale Giulio Pastore, 6 – 00144 Roma

ISBN 978-88-7484-301-5

Stampato presso la Tipolitografia INAIL – Milano, maggio 2013

# Infortuni sul lavoro. Un modello di lettura (della numerosità) su "open data" dell'Inail

Giacomo Ciriello, Massimo De Felice, Rosaria Mosca, Massimiliano Veltroni

**Sommario** – In questo *Quaderno* si propone e si argomenta il principio per cui non ha senso pratico parlare di "open data" se non si definisce e si rende esplicito il "modello di lettura" dei dati. Vengono individuate le componenti strutturali del modello (elenco delle grandezze da rilevare, i loro significati e interrelazioni: definiti in un vocabolario e strutturati con un thesaurus); sono considerati gli strumenti di tutela dell'affidabilità dei risultati (processi di "data quality", politica dei dati, qualità degli algoritmi).
Il principio del "modello di lettura" è applicato all'analisi della numerosità degli infortuni sul lavoro. Il modello è definito su un insieme di 58 tabelle, organizzate "a strati", per livello di approfondimento. Sono illustrati percorsi di analisi orizzontali, sullo stesso strato; linee di approfondimento tematico (a partire dai temi di livello zero, nella logica del "drill down"); letture trasversali, guidate dalle classi di raggruppamento dei dati.
L'insieme delle tabelle è proposto come corredo dei dati elementari necessari per generarlo. I dati elementari (con la struttura dei metadati), l'insieme delle tabelle, il vocabolario e il thesaurus costituiscono il blocco di informazioni che l'Inail potrebbe rendere pubblico (in ottemperanza ai principî dell'"open data"). La costruzione del blocco delle informazioni (illustrata in questo *Quaderno*) è stata realizzata con un processo informatico, in forma di prototipo, che potrà garantire la periodicità delle pubblicazioni. Il processo informatico è predisposto per sostenere ampliamenti della base informativa (sulle malattie professionali e con i dati del "settore navigazione"). Negli allegati è riportato un esempio di "output" del prototipo informatico.

*Parole chiave, frasi*: infortuni sul lavoro, infortuni con esito mortale, open data.

Gli autori svolgono o hanno svolto la loro attività presso l'Inail: Giacomo Ciriello è stato Capo della Segreteria tecnica di presidenza, Massimo De Felice come Presidente; Rosaria Mosca nella Direzione centrale Servizi informativi e telecomunicazioni, Massimiliano Veltroni nella Consulenza statistico-attuariale. Si ringraziano la tecnostruttura dell'Inail – in particolare la Direzione centrale Servizi informativi e telecomunicazioni, la Consulenza statistico-attuariale e l'Avvocatura – per la collaborazione; Orietta Amato, Patrizio Rossi, Gabriella Salinetti per le discussioni sui temi di questo *Quaderno*; l'UNI (l'Ente Nazionale Italiano di Unificazione) per l'ausilio sui temi della normazione. Alcune parti di questo *Quaderno* sono state presentate al seminario su "L'open data per lavoro e previdenza", organizzato dal Ministero del Lavoro e delle Politiche Sociali (Roma, 14 febbraio 2013); in una riunione del Consiglio di Indirizzo e vigilanza dell'Inail (il 27 febbraio 2013); nella "riunione con le parti sociali" convocata dal Ministro del Lavoro e delle Politiche Sociali (Roma, 13 marzo 2013): in tutte queste occasioni si sono avuti commenti e suggerimenti utili. Gli autori hanno la piena responsabilità delle opinioni espresse (che non rappresentano necessariamente la posizione ufficiale dell'Inail) e degli eventuali errori.
Commenti possono essere inviati a m.veltroni@inail.it.

## Indice degli argomenti

Premessa 7

1 Gli "open data", l'importanza del "modello di lettura" dei dati 8

2 Rendere informativi i "dati aperti" 9

3 Vocaboli e frasi del "modello di lettura" (della numerosità degli infortuni) 12
3.1 L'infortunio nel linguaggio comune e nel linguaggio speciale (la definizione operazionale) 12
3.2 Il nucleo del linguaggio speciale 13
3.3 Il thesaurus 24

4 Su alcuni processi amministrativi rilevanti per l'interpretazione dei dati 25

5 "Viste" sui dati, linee di approfondimento, altri itinerari di lettura 27
5.1 Caratteristiche temporali dei dati 27
5.2 Sui casi "negativi" d'infortunio con esito mortale 28
5.3 Le "viste" 30
5.4 Leggere i dati 31

6 Dati (aperti) per replicare gli itinerari di lettura con diverso dettaglio 32
6.1 I dati elementari per la lettura annuale 32
6.2 I dati elementari per la lettura "di periodo" 33
6.3 Questioni di privacy 33
6.4 Un calendario e la forma di pubblicazione dei dati 34

7 Postilla storica 35

Riferimenti bibliografici 38

Allegato 1
La lettura della numerosità degli infortuni, sugli "open data" dell'Inail (anno 2011) 41

Allegato 2
La numerosità degli infortuni sui dati "di periodo" (aprile 2011) 99

***Premessa*** – Questo *Quaderno* raccoglie i risultati della prima fase di un progetto di ricerca finalizzato a individuare un "nucleo di dati" elementari, da rendere pubblico (nel senso dell'"open data"), che possa rappresentare in modo adeguato i fenomeni degli infortuni sul lavoro e delle malattie professionali. La rappresentazione dei fenomeni potrà dare sostegno ulteriore alla definizione e alla verifica delle politiche di prevenzione, per garantire la sicurezza nei luoghi di lavoro.
La prima fase del progetto ha riguardato il disegno di un "modello di lettura" dei dati sugli infortuni, raccolti dall'Inail (nello svolgimento delle proprie attività istituzionali). Il modello è finalizzato all'analisi quantitativa del fenomeno (numero di infortuni per data di accadimento, per classi degli infortunati e caratteristiche dell'infortunio); è strutturato in "viste", realizzate con chiavi di selezione e criteri di raggruppamento; le viste sono rappresentate in 58 tabelle, che possono servire a comporre diversi itinerari tematici, a guidare diversi livelli di approfondimento sui dati elementari "aperti" (con elaborazioni autonome da parte degli "utenti" [1]).
I principî definiti per la costruzione del "modello di lettura", le componenti del modello (dati, informazioni di corredo: metadati, vocabolario, thesaurus, viste) e la struttura del processo informatico realizzato per produrre gli "open data" degli infortuni sono generali, e valgono per l'intero progetto.
Nelle fasi successive del progetto il "modello di lettura" (e quindi l'insieme dei dati elementari e delle informazioni di corredo) verrà esteso a considerare le malattie professionali, il "settore navigazione" (per infortuni e malattie), la dimensione economica dei fenomeni (costo degli infortuni e delle malattie, per l'Inail e per le imprese), e a definire e misurare un "indice di sinistrosità", che consenta di valutare il numero di infortuni in rapporto al periodo di esposizione al rischio e al livello di rischiosità dell'esposizione.

Questo *Quaderno* è composto da 7 capitoli, e 2 allegati. Nel primo capitolo si pone il tema dell' "open data" nella pubblica amministrazione; si propone e si argomenta il principio per cui non ha senso pratico parlare di "open data" se non si definisce e si rende esplicito il modo di rappresentazione dei fenomeni cui i dati si riferiscono (il "modello di lettura" dei dati). Il capitolo 2 delinea le modalità per rendere informativi i "dati aperti" dell'Inail: l'importanza del vocabolario (del linguaggio speciale, dei termini tecnici e di gergo), le potenzialità del thesaurus (per verificare la coerenza logica dei lemmi del vocabolario), la rilevanza della politica dei dati a tutela della "data quality", l'attenzione alla "interoperability" (per il collegamento col "Sistema informativo per la prevenzione nei luoghi di lavoro", e con il progetto dell'"European Statistics on Accidents at Work"). Il capitolo 3 definisce il nucleo del linguaggio speciale per leggere il fenomeno degli infortuni sul lavoro (con i 69 lemmi del vocabolario, e le 17 catene di significati del thesaurus). Nel capitolo 4 si danno le precisazioni su alcuni processi amministrativi rilevanti per l'interpretazione dei dati. Il capitolo 5 definisce il "modello di lettura" del numero di infortuni: i criteri statistici di rilevazione dei dati e delle loro qualificazioni; la forma delle rappresentazioni (in 58 tabelle, organizzate "su temi" e "a strati", per livello di approfondimento); le modalità di applicazione del modello (per "viste" annuali, o su frazioni d'anno). Nel capitolo 6 è fornito e decritto l'elenco dei dati

[1] Il "Codice in materia di protezione dei dati personali" (DLgs 196/2003 e s.m.i.) definisce "utente, qualsiasi persona fisica che utilizza un servizio di comunicazione elettronica accessibile al pubblico, per motivi privati o commerciali, senza esservi necessariamente abbonata" (art. 4, comma 1, lettera g).

da rendere "open", per consentire agli utenti di replicare gli itinerari di lettura, con diverso dettaglio. La postilla storica (capitolo 7) documenta le esperienze che sono state utile riferimento per definire il "modello di lettura".
Negli allegati è riportato un esempio di "output" del prototipo informatico costruito per applicare il modello ai dati e produrre l'insieme delle tabelle (annuali nell'allegato1, di periodo nell'allegato 2); l'"output" ha finalità soltanto metodologica: è proposto per dare concretezza agli schemi di lettura (grandezze selezionate, criteri di classamento).

## 1 Gli "open data", l'importanza del "modello di lettura" dei dati

*Tipi di dato della pubblica amministrazione* – La pubblica amministrazione produce in generale due tipi di dato: 1 – dati gestionali, relativi ai processi organizzativi e di servizio; 2 – dati di tipo statistico, descrittivi dei fenomeni che gestisce. Sebbene sia affermato e condiviso il giudizio sull'importanza dell'"adeguata diffusione" e "trasparenza" (secondo i principî dell'"open data") di queste informazioni, le norme richiamano il riferimento più esplicito ai dati gestionali (di tipo 1)[2].
Per garantire l'efficacia della diffusione è importante considerare che la pubblicazione dei dati implica livelli di complessità molto diversi tra i due tipi: i dati di tipo statistico (poiché rilevati con primaria finalità amministrativa) richiedono particolari accortezze, per aver garantita la qualità di "statistiche ufficiali".

*Il modello di lettura* – Per i dati di tipo statistico, non ha senso pratico parlare di "open data" se non si definisce e si rende esplicito (e quindi trasparente) il "modello di lettura" dei dati[3].
Definire il modello significa esprimere un'opinione sui modi di rappresentazione del fenomeno (i fatti), cui i dati si riferiscono: quali sono le grandezze ultime necessarie per descrivere in modo adeguato il fenomeno; e poi, come organizzarle in tabelle per agevolarne il confronto, quali sono le definizioni non ambigue ("operazionali") delle grandezze, quali criteri sono utilizzati per rilevarle o quali formule (algoritmi) servono per calcolarle, quali dati elementari servono per "alimentare" le formule (e realizzare le tabelle). Con questo processo si identifica l'insieme dei dati neces-

---

[2] Si dice che "l'adeguata diffusione di queste informazioni (dati statistici e territoriali, rapporti socio-economici e ambientali, normativa, ecc.) può rappresentare un importante elemento per favorire la crescita economica e produttiva, la ricerca, l'innovazione, la competitività" [DaI-011, 49]. Ma l'articolo 11 del DLgs 150/2009 definisce la "trasparenza" come "accessibilità totale, anche attraverso lo strumento della pubblicazione sui siti istituzionali delle amministrazioni pubbliche, delle informazioni concernenti ogni aspetto dell'organizzazione, degli indicatori relativi agli andamenti gestionali e all'utilizzazione delle risorse per il perseguimento delle funzioni istituzionali, dei risultati dell'attività di misurazione e valutazione svolta dagli organi competenti, allo scopo di favorire forme diffuse di controllo del rispetto dei principi di buon andamento e imparzialità. Essa costituisce il livello essenziale delle prestazioni erogate dalle amministrazioni pubbliche ai sensi dell'art. 117, secondo comma, lettera m, della Costituzione". I "contenuti minimi dei siti web istituzionali" (corredati da norma di riferimento, eventuali vincoli, indicazioni di reperibilità) individuano soltanto dati di tipo gestionale [DaI-011, 33-38].

[3] Il riferimento al "modello di lettura" richiama, in grande, il problema fondamentale dell'inferenza statistica. Il senso del problema è stato posto (da un grande probabilista) argomentando su una frase di Pirandello: "un fatto è come un sacco; vuoto, non si regge"; per giudicare i fenomeni (e utilizzare i dati che li descrivono) serve altro: "occorre *in più* anche una qualche opinione per interpretarli e per utilizzarli sensatamente" [dF-971, 97]. D'altra parte il dibattito recente sulle grandi quantità di dati da lavorare con l'"agnostic statistics" (sintetizzato in Anderson, C., *The end of theory: will the data deluge makes the scientific method obsolete?*, http://www.edge.org, 2008) è solo una rilettura modernista di vecchi dialoghi; e l'idea di una "epistemologia con la forza bruta" pone la domanda: con una ricerca guidata da nessuna ipotesi teorica, non ispirata da alcuna intuizione, "che conoscenza possiamo ottenere?" (Casati, R., *Prima lezione di filosofia*, Bari, Laterza, 2011, 173).