La qualità dei dati: concetti e misure

di Domenico Natale

Si parla sempre più spesso di qualità dei dati, mentre si susseguono pubblicazioni sul tema anche da parte dell'UNI/ISO, come l'UNI ISO/IEC 25012:2014 "Modello di qualità dei dati" e l'ISO/IEC 25024:2015 "*Measurement of data quality*". Il tema richiama i termini arcaici di EDP - Electronic Data Processing e di CED - Centro Elaborazione Dati, e si assiste nei data center di oggi a una crescita inarrestabile di dati. Di pari passo tra le professioni si affiancano al database administrator, il data architect e il data scientist.

I dati sembrano un argomento nuovo, mentre sono stati da sempre alla base dell'Informatica (informazione automatica) e della statistica. A rafforzare la tendenza di maggiore attenzione ai dati contribuiscono i termini attualmente in voga: big data, data cloud, open data.

Gli standard sulla qualità dei dati sono parte integrante del progetto SQuaRE, definito nell'ISO/IEC 25000:2014 "S*ystems and software Engineering - Systems and Software Quality Requirements and Evaluation – Guide to SQuaRE*". Il progetto, come evoluzione dell'ISO 9126 ormai obsoleto, pone attenzione, tra l'altro, al dato come componente integrata del sistema, del software e dei servizi, definendo appositi standard concettuali e metrici.

Gestire la qualità dei dati è importante per varie ragioni, ad esempio:
- favorire il governo della crescente disponibilità di dati;
- acquisire dati la cui qualità è sconosciuta;
- gestire informazioni spesso insoddisfacenti;
- contenere la dispersione di dati nei sistemi e tra sistemi;
- incrementare i dati riusabili;
- far convivere i legacy system con sistemi aperti;
- ridurre duplicazioni di dati e impegno risorse;
- diminuire i costi della non qualità;
- eliminare progressivamente modelli cartacei e processi manuali;
- rimuovere le cause che determinano dati errati;
- contribuire allo sviluppo di servizi web innovativi.

La qualità dei dati (e delle informazioni) sta divenendo una componente essenziale della qualità in uso dei servizi IT intesa come efficacia, efficienza, soddisfazione, copertura del contesto, mitigazione dei rischi economici, ambientali e della salute.
Ma prima di approfondire il tema occorre far riferimento al ciclo di vita del dato (definito nell'ISO 25024) composto di fasi che vanno dalla prima definizione di dato alla sua cancellazione. Per dato (definito nell'UNI 25012) si intende la "rappresentazione re-interpretabile di informazioni in un modo formalizzato utilizzabile per la comunicazione, l'interpretazione o l'elaborazione".
Esistono informazioni che nascono in natura, altre che riguardano informazioni create dall'uomo, che si possono quindi definire artificiali o convenzionali. Tra le numerosissime informazioni che nascono in natura si possono considerare quelle presenti ad esempio in un raggio di luce (i dati dello spettro, forse i più antichi), nel carotaggio di ghiaccio artico (i dati atmosferici di secoli fa), in un tronco d'albero (l'età e le condizioni climatiche), i dati presenti nel corpo umano ecc.. Sono questi i dati, incorporati nella sostanza oggetto di ricerca, studiati da parte di fisici, biologi, matematici, medici, ecc.. Essi ci rinviano al significato originario di dato, cioè "donato", che si impone all'esperienza umana e che diventa l'oggetto di studio e analisi di fenomeni. Tali dati si ritengono "veri" per definizione e la sua determinazione è il risultato della ricerca scientifica, che si affida agli strumenti utilizzabili sempre più sofisticati e all'analisi teorica.

Tra i dati creati dall'uomo si possono invece considerare tutti quelli che rappresentano informazioni digitali convenzionali riportate nei *MIS - Management Information System*: sono in genere dati con formati strutturati all'interno di un sistema informatico. Essi attraversano tutte le fasi del ciclo di vita dei dati: dal design, all'integrazione, all'elaborazione, alla cancellazione, ecc..Per il loro corretto trattamento si rivela molto utile, a priori con un approccio forward engineering, attenersi ai modelli di qualità per le fasi di definizione dei requisiti. Tali modelli, seppur dedicati ai dati strutturati, tengono conto di tutti i tipi di dati (ad esempio: stringhe di caratteri, testi, date, numeri, immagini, suoni), gestiti attraverso svariati mezzi (carta, computer, dispositivi mobili, ecc.). A posteriori, quando i dati sono disponibili, i modelli si rilevano utili per valutazioni sistematiche della qualità; nei casi in cui la qualità è ridotta si parla di Data cleaning, per depurare i dati, e nei casi complessi di data mining per scoprire informazioni "nascoste". A seconda dei casi la valutazione, accompagnata da attività di benchmark, può indurre ad attività di re-engineering.
Per i dati non strutturati si apre anche una nuova strada di applicazione che fa ricorso alle tecniche statistiche dei big data che si rivelano particolarmente utili per le sorgenti di dati, varie e vaste. Il ricorso alle tecniche dei big data, specie nelle fasi di integrazione e elaborazione, tendono a enucleare "il costante nel variabile" traendo informazioni da dati apparentemente caotici e non correlati.
In entrambi i casi, dati strutturati o meno, nelle fasi di memorizzazione sta emergendo l'approccio del data cloud e, nelle fasi di presentazione e visualizzazione, l'approccio open data che promuove la diffusione dei dati in formati aperti per garantire trasparenza e agevolare nuove ricerche, nuove partecipazioni degli utenti all'interpretazione e maggior uso dei dati.

TABELLA 1 - FASI DEL CVD - CICLO DI VITA DEI DATI E TECNICHE RILEVANTI

CVD ISO/IEC 25024	Misure definite nel 25024 per dati strutturati	Big data	Data cloud	Open data
Design, acquisizione, raccolta dati	Dati relativi a architetture, modelli, documenti	Nuove modalità di raccolta dati, anche in tempo reale		
Integrazione, elaborazione, memorizzazione, cancellazione	Dati relativi ai file	Dati strutturati e destrutturati provenienti da fonti varie e di grandi dimensioni, con qualità non sempre definita (dati naturali e artificiali)	Dati strutturati e destrutturati virtualizzati e anche affidati a terzi	Elaborazioni in formati aperti
Presentazione	Dati relativi a contenuti visualizzati	Infografica		Dati pubblici di interesse economico, sociale

I dati strutturati di elevata qualità potranno migrare verso i big data, il data cloud e l'open data. E' prevedibile che il controllo di qualità potrà attuarsi con processi che si adatteranno dinamicamente a misurazioni anche in tempo reale.
Le applicazioni con dati duplicati andranno a ridursi e le organizzazioni, supportate da customer relationship management, avvicineranno dati "in ritardo" con dati sempre più "accurati e tempestivi", generando diffusi miglioramenti dei servizi. Si intuisce che tali approcci non sono esclusivi, ma andranno ad armonizzarsi e convergere, al crescere anche delle infrastrutture connesse.
Nella presente analisi ci si limiterà ad approfondire il possibile uso dei modelli di qualità dei dati, definiti nell'ISO 25012 e 25024, esaminandone brevemente i concetti essenziali con alcuni esempi di misure, consapevoli che nel prossimo futuro le fasi classiche del ciclo di vita del dato non potranno più prescindere dalle altre tecniche citate che si stanno affermando sul mercato.

ISO/IEC 25012 Modello di qualità dei dati

Si riporta di seguito, in sintesi, il modello di qualità dei dati definito nello standard internazionale ISO/IEC 25012 del 2008, divenuto norma italiana nel 2014 con la sigla UNI ISO/IEC 25012. Il modello fornisce un elenco di quindici caratteristiche considerate da due punti di vista: inerenti i dati e dipendenti dal sistema.
Per inerente si intende il caso in cui la qualità del dato si riferisce alle sue proprietà intrinseche, a prescindere dal supporto di rappresentazione utilizzato e dagli aspetti tecnologici. Per dipendente dal sistema si intende il caso in cui la qualità del dato è influenzata dal sistema o strumento che lo ospita. Alcune caratteristiche sono prevalentemente interessate da un solo punto di vista, altre da entrambi:

Inerente

- accuratezza, intesa come perfetta rispondenza del dato con la realtà che rappresenta;
- attualità, cioè il giusto tempo con il quale il dato è creato o aggiornato;
- coerenza, dato non contraddittorio con altri dati, all'interno del sistema o tra sistemi;
- completezza, ove tutti gli attributi necessari sono presenti, con tutte le fonti;
- credibilità, nel caso in cui la fonte del dato è certa.

Inerente e dipendente dal sistema

- accessibilità, il dato è accessibile in varie situazioni, anche da disabili;
- comprensibilità, il significato del dato (e del metadato) è chiaro e immediato;
- conformità, il dato risponde prioritariamente a regolamentazioni, anche locali;
- efficienza, il dato è gestito con risorse accettabile e tempi adeguati allo scopo;
- precisione, il dato possiede il livello di misura discriminante necessario;
- riservatezza, il dato può essere utilizzato solo da utenti autorizzati;
- tracciabilità, gli accessi al dato sono registrati.

Dipendente dal sistema

- disponibilità, il dato è disponibile e interrogabile;
- portabilità, il dato è gestibile e migrabile in diversi ambienti operativi;
- ripristinabilità, il dato è salvato in un ambiente sicuro e recuperabile.

Molti dei termini o concetti di uso comune possono essere connessi, per chiarezza, al modello proposto.

TABELLA 2 - CARATTERISTICHE DEI DATI UNI ISO/IEC 25012 E SINONIMI

Caratteristiche	Dimensioni, fattori, proprietà
Accuratezza	Correttezza, affidabilità, veridicità, realtà, sintassi-semantica, validità, attendibilità, oggettività
Attualità	Tempestività, giusto tempo, tempo reale, aggiornamento
Coerenza	Allineamento, consistenza, qualità cross-enterprise
Completezza	Volume, saturazione, copertura attributi
Credibilità	Ufficialità o fonte autorizzata
Accessibilità	Inclusione, multicanalità, disabilità, documenti o contenuti accessibili
Comprensibilità	Intuitività, leggibilità, riconoscibilità, interpretabilità
Conformità	Adeguatezza a norme o standard nazionali, convenzioni, regolamenti
Efficienza	Gestione e impegno tempo e risorse, occupazione spazio
Precisione	Grado di dettaglio, esattezza, grado di approssimazione, elemento discriminante
Tracciabilità	Log, memorizzazione e storicità di accessi
Riservatezza	Confidenzialità, sicurezza, segretezza, cifratura, privacy, protezione, autorizzazione
Disponibilità	Continuità del servizio, finestre di servizio, orari, trasparenza
Portabilità	Migrabilità, adattamento, installabilità, multi-device, interscambio, formato standard
Ripristinabilità	Ricoverabilità, copie di backup, disaster recovery, salvataggio

L'adesione a un modello comune di qualità è una precondizione per l'integrazione dei dati e la realizzazione di banche dati dialoganti, la conformità alle direttive normative/legislative, la diffusione di una cultura inter-amministrativa, il miglioramento e semplificazione dei procedimenti burocratici secondo approcci condivisi.

ISO/IEC 25024 Misurazione della qualità dei dati

Si riporta di seguito una breve sintesi dello standard ISO/IEC 25024 di fine 2015 sulla misurazione qualità dei dati che estende alla misurazione lo standard concettuale ISO/IEC 25012.

Lo standard definisce 63 misure, ed è articolato nel seguenti capitoli:

- Ambito di applicazione;
- Definizione dei termini;
- Misure basilari di riferimento:
  - sigla identificativa della misura;
  - breve descrizione;
  - funzione o algoritmo di calcolo;
  - fase del ciclo di vita del dato a cui si applica la misura;
  - prodotto specifico di applicazione;
  - proprietà;
  - allegati informativi.

Lo standard non prescrive soglie di accettazione di qualità, che saranno stabilite dall'utilizzatore nello specifico contesto d'uso. Esso suggerisce però esempi di rilevanza delle misure e spiegazione dei casi più complessi. Le misure di qualità dei dati proposte nello standard, in lingua inglese e a cui si rinvia per approfondimenti, non sono esaustive e il metodo consente facili ampliamenti, secondo le esigenze nel contesto di applicazione. La maggioranza delle misure si riferisce ai file, compresi i data base relazionali, ai modelli di dati, a elementi dell'architettura, documenti, modulistica e device di presentazione (vedi Grafico 1).

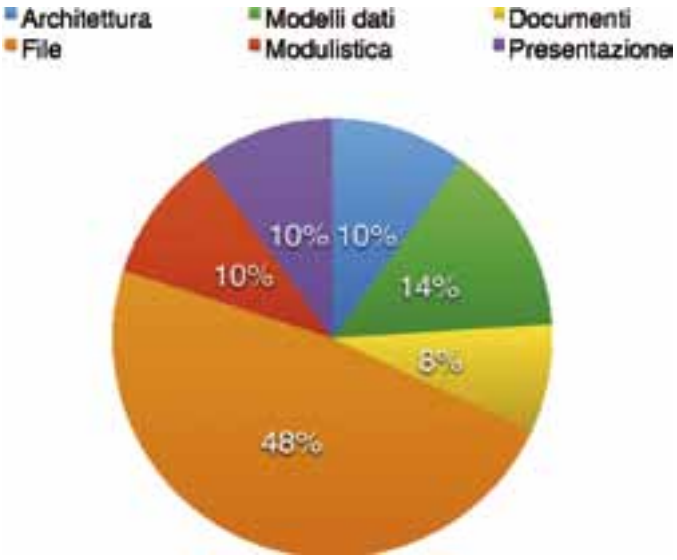

Grafico 1 - Ripartizione misure per prodotto

A titolo esemplificativo alcuni elementi delle misure di qualità, scelti tra quelli disponibili, sono di seguito riportati; nella pratica tali elementi possono essere aggiornati, dandone motivazione, sulla base delle esigenze dell'utente e del contesto d'uso:

- accuratezza: outlier (rischio di inaccuratezza dei valori quantitativi di una distribuzione, fortemente fuori media) da verificare con ispezione e eventualmente da confermare;
- attualità: informazioni per le quali è prescritta la periodicità di aggiornamento;
- coerenza: formato della stessa informazione in diversi file; informazioni duplicate (rischio di incoerenza);
- completezza: attributi definiti nei modelli dati; attributi con metadati definiti nel dizionario dati;
- credibilità: informazioni validate con un processo specifico;
- accessibilità: dati (o documenti) accessibili da utenti specifici (es. disabili);

- comprensibilità: valori rappresentati da simboli conosciuti;
- conformità: dati (o documenti) conformi a standard o regolamenti;
- efficienza: dati memorizzati in formati qualificati come efficienti; valori ritenuti facilmente usabili dagli utenti; spazio occupato da eventuali duplicazioni; tempo di ritardo di aggiornamento di un dato tra diversi sistemi;
- precisione: campi riportati con la precisione richiesta utile per discriminare;
- riservatezza: cifratura e decrittazione delle informazioni;
- tracciabilità: valori con tracciabilità degli accessi;
- disponibilità: dati disponibili in un dato periodo di tempo;
- portabilità: dati che preservano un certo livello di qualità dopo la migrazione;
- ripristinabilità: dati ricoverati e recuperati dal sistema con successo e correttezza.

Considerazione conclusiva

L'adozione dei modelli di qualità, integrati con i requisiti del sistema, del software, dei dati e dei servizi, agevola la misurazione della qualità raggiunta, la valutazione dei risultati, le attività di *re-engineering* e *benchmarking,* consentendo eventuali aggiustamenti iterativi dei requisiti (vedi grafico 2).
La fase di valutazione, supportata dal rapporto di conformità, consente anche di fornire *feedback* all'organizzazione e di porre le basi per eventuali successive attività di pre-certificazione.

Domenico Natale
Presidente Commissione UNINFO SC7 Ingegneria del Software
Editor delle norme ISO/IEC 25012:2008, UNI ISO/IEC 25012:2014 e ISO/IEC 25024:2015

THE QUALITY OF DATA: CONCEPTS AND MEASURES

The quality of data is more and more getting a concern today and more publications on this topic are being issued also by UNI/ISO, such as ISO/IEC 25012: 2014 "Model data quality" and ISO/IEC 25024: 2015 "Measurement of data quality." This subject calls back to the archaic terms of EDP - Electronic Data Processing and DPC - Data Processing Centre and the number of data is continously growing in today's data centers, while the professional role of data architects and data scientists is being aligned with that of database administrator.

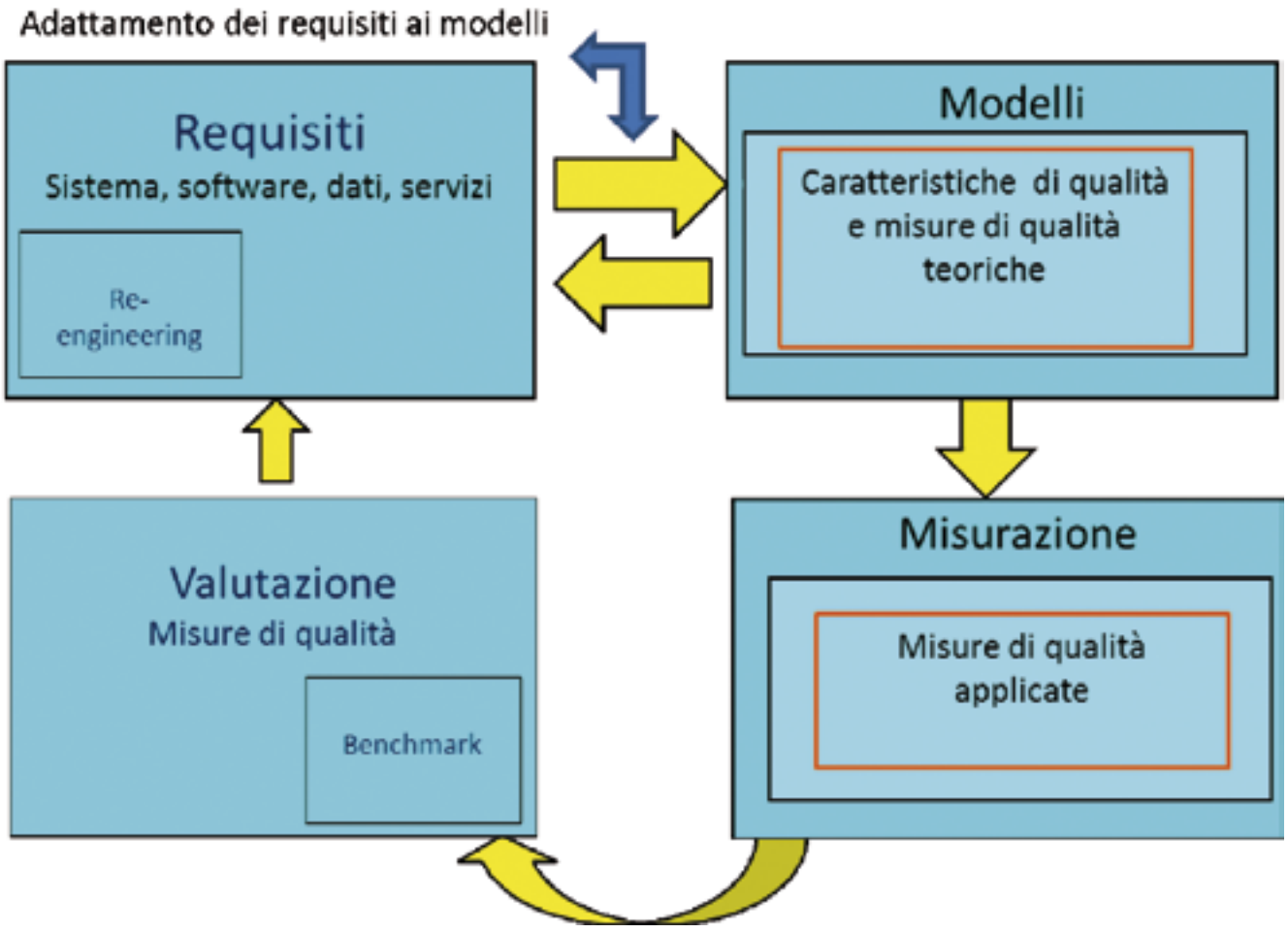

Grafico 2 - Ciclo della qualità basato su ISO 25000 - SQuaRE