

Qualità dei dati e Standard ISO/IEC 9126: Analisi Critica ed Esperienze nella Pubblica Amministrazione Italiana

Domenico Natale¹, Monica Scannapieco², Paola Angeletti¹, Tiziana Catarci², Gianluigi Raiss³

¹ SOGEI-Società Generale di Informatica

² DIS-Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza"

³ AIPA-Autorità per l'Informatica nella Pubblica Amministrazione

1. Introduzione

I governi dei principali paesi industrializzati stanno sempre più decisamente puntando su strategie di *e-government*, finalizzate a raggiungere in tempi relativamente brevi due principali obiettivi: migliorare la qualità dei servizi resi ai cittadini-utenti, aumentare la efficienza ed efficacia delle strutture pubbliche. Ambedue questi obiettivi si fondano su un approccio del tutto nuovo alle tecnologie informatiche, reso possibile dallo sviluppo di internet e degli strumenti di comunicazione. Un approccio che ha come immediata conseguenza il porre l'utente, e non più gli atti amministrativi, come soggetto fondamentale ed attivo del processo di erogazione dei servizi.

Il presupposto comune a queste iniziative è la qualità della informazione, scambiata, elaborata, utilizzata nelle transazioni tra cittadino e Pubblica Amministrazione ma anche tra i diversi Enti della Pubblica Amministrazione. La qualità di questa informazione è perciò uno dei fattori abilitanti della innovazione e del miglioramento dei servizi. Basta pensare a quanti servizi resi ai cittadini si basano su elementi ricavati dai dati anagrafici, o su dati fiscali, o come le statistiche ricavate da questi dati influenzano le politiche di governo e l'azione amministrativa. Ma oltre a condizionare l'efficacia complessiva dell'azione della Pubblica Amministrazione, ed a deteriorare l'immagine presso gli utenti, la cattiva qualità dell'informazione aumenta i costi di erogazione dei servizi e quelli a carico dei cittadini, rendendo difficile anche ogni intervento di re-engineering dei procedimenti e la definizione delle strategie a medio termine.

In definitiva, sono tutti d'accordo che la scarsa qualità delle informazioni risulta sempre più pervasiva nella nostra Società soprattutto in quello schema di architettura a rete verso cui tende il sistema Pubblica Amministrazione.

Tuttavia, migliorare la qualità dei dati nel contesto della Pubblica Amministrazione è un processo complesso, ed oneroso, per diversi motivi:

- la grande quantità di informazioni gestite dalla Pubblica Amministrazione, e la molteplicità dei soggetti che ne hanno la responsabilità a diversi livelli,
- la dispersione dei dati presso diversi soggetti, secondo logiche quasi sempre funzionali alle esigenze delle singole organizzazioni, senza una visione coerente, integrata e trasversale finalizzata alla interoperabilità ed alla cooperazione,
- la disomogeneità in particolare, ogni fonte amministrativa che nel tempo ha raccolto e gestito archivi, ha usato concetti, definizioni, classificazioni, regole di iscrizione e cancellazione proprie, vincolate a funzioni ed adempimenti specifici,
- la persistenza, nelle diverse organizzazioni, di numerose categorie di dati, creati ed elaborati internamente, che si potrebbero definire "specialistici", i quali, pur essendo di interesse anche per altre organizzazioni, non sono immediatamente riutilizzabili, a causa della loro ambiguità semantica, o della loro solo parziale acquisizione in forma elettronica, o del mancato controllo di congruità con gli altri dati correlati esistenti,

- la frequente necessità di far convivere, nella stessa organizzazione, architetture e sistemi legacy e sistemi distribuiti, spesso con differenti basi di dati, con problemi di replicazione, ridondanza, allineamento, coerenza,
- la pressoché totale assenza di procedure ad hoc di gestione della qualità dei dati, ma, spesso, anche di procedure di gestione dei dati autonome rispetto alle funzionalità offerte dai sistemi.

Sono problemi complessi da risolvere, per superare i quali occorre investire sia in un'opera di pulizia dei dati, e di riorganizzazione/razionalizzazione delle loro architetture, che ne permetta un utilizzo efficace e trasversale ai processi amministrativi ed ai servizi che coinvolgono più amministrazioni, sia nel costruire adeguati processi di gestione dei dati e della loro qualità svincolati dalle funzionalità applicative e condivisi e diffusi nell'intero sistema Pubblica Amministrazione.

Obiettivi di questi processi dovrebbero essere non solo evitare la ridondanza e la replicazione fuori controllo, ma anche garantire l'aggiornamento dei dati in modo tempestivo, gestire la sicurezza degli accessi, facilitare la manutenzione dei dati. Una organizzazione funzionale a tali obiettivi prevede la separazione degli strati logici del sistema informativo relativi all'analisi e progettazione delle basi di dati, dell'accesso ai dati (lettura, aggiornamento, cancellazione, monitoraggio della qualità pulizia), dell'elaborazione dei dati al fine del loro utilizzo.

Dal punto di vista metodologico, l'attuazione di questa organizzazione si basa, principalmente, sul preliminare censimento delle basi di dati esistenti e delle procedure che le gestiscono (e della documentazione relativa agli schemi logici dei data base e dei processi), sulla descrizione dei flussi in ingresso ed uscita ed all'interno dei sistemi (con l'ottica di vedere la pubblica amministrazione, per quanto possibile, come un unico sistema nei confronti del cittadino-utente), sulla individuazione dei vincoli di qualità che devono rispettare i dati. Il rispetto di questi vincoli deve quindi essere misurato, per poter valutare lo stato dell'arte ed impostare eventuali azioni correttive e migliorative.

Tutto ciò è però possibile se si dispone di adeguati indicatori della qualità dei dati, che siano di facile comprensione ed utilizzo e siano inseriti all'interno di un contesto di larga diffusione e di un modello delle loro relazioni con le altre componenti della qualità dei servizi resi agli utenti. Perciò, questi indicatori devono essere integrati con il modello di qualità delle applicazioni che fanno uso dei dati, e dei processi che utilizzano e trasformano i dati stessi. Misurando questi indicatori ed analizzando le misure rilevate, si devono quindi poter attivare, a 360 gradi, le eventuali azioni volte a migliorare la qualità dei dati non rispondenti ai requisiti, nell'ottica di fornire comunque ai cittadini un servizio pari alle loro esigenze ed aspettative.

Per quanto detto, il riferimento ad un modello di qualità consolidato e diffuso, come quello definito dalla norma ISO/IEC 9126, adattato alla specificità del dato piuttosto che del prodotto software, sembra essere la soluzione migliore per costruire un framework di caratteristiche della qualità dei dati. Questo modello definisce infatti un insieme di caratteristiche e sottocaratteristiche di qualità tra loro correlate, in grado di evidenziare aspetti intrinseci all'entità in esame, aspetti esterni, legati all'ambiente di utilizzo, ed aspetti di qualità nell'uso che dell'entità fa un utente. Coprendo così, da diversi punti di vista, le esigenze di qualità che possono avere i diversi soggetti interessati al dato. Nel seguito, viene quindi proposto un primo adattamento delle caratteristiche ISO/IEC 9126 alle specificità della qualità dei dati.

Impostare una metodologia e definire degli indicatori, od anche un modello di qualità non è però sufficiente a risolvere i problemi sopra individuati. La Pubblica Amministrazione, nel suo insieme, deve definire una politica comune nei confronti dell'informazione, deve aumentare la interoperabilità e la cooperazione al suo interno, deve accelerare l'innovazione tecnologica ed organizzativa, deve ampliare le possibilità di comunicazione con i cittadini.

Diverse iniziative sono state avviate per accelerare l'attuazione di questo programma, tra cui la razionalizzazione degli archivi anagrafici del Ministero dell'Economia e delle Finanze, il progetto "qualità del dato" in INAIL, il repertorio integrato degli Agenti Economici, che vede coinvolti ancora l'amministrazione finanziaria, INPS, le Camere di Commercio, le esperienze presso il Ministero del Tesoro per la definizione di un data warehouse di supporto agli acquisiti nella Pubblica Amministrazione. Questo articolo è un contributo in questa direzione, per offrire un modello di confronto ed una terminologia comune alle diverse iniziative in corso.

2. Definizioni

Si sceglie di assumere come definizione del concetto di “qualità”, quella fornita dallo standard ISO 8402 “Quality Management and Quality Assurance-Vocabulary” [7], riportata anche nella ISO/IEC 9126 “Information Technology- Software Product Evaluation- Quality Characteristics and Guidelines for their Use” [8].

Nello standard 8402 la qualità è genericamente definita come segue:

“l’insieme delle caratteristiche di un’entità che conferiscono ad essa la capacità di soddisfare esigenze espresse ed implicite”

In base a tale definizione si possono evidenziare le seguenti due proprietà

- la qualità di un’entità è un concetto soggettivo, fortemente dipendente dai requisiti che l’utente dell’entità richiede in maniera implicita o esplicita;
- la qualità è rapportata ad un insieme di caratteristiche, è dunque un concetto multidimensionale, ovvero determinabile in base a diverse componenti piuttosto che in base ad una singola.

In questo lavoro, si propone di specializzare la definizione di qualità della ISO 8402 al caso specifico della *qualità dei dati*.

Infatti, la qualità dei dati gode di entrambe le proprietà descritte, ed in particolare:

- il carattere soggettivo è esplicito nelle diverse definizioni che sono presentate in letteratura, tra cui si citano come esempio “conformance to requirements” [5] e “fitness for use by information consumers” [14];
- in [12] sono presentate 17 dimensioni associate alla qualità dei dati, classificabili in base alla vista concettuale dei dati, ai valori ed al formato degli stessi. Anche in [14] è proposto un insieme di 17 dimensioni, differente da quello proposto in [12], e raggruppabile in categorie, rispetto alle caratteristiche di dipendenza intrinseca, accessibilità, dipendenza dal contesto, rappresentazione.

Sulla base di tali considerazioni, si è scelto di adottare come definizione di qualità dei dati la definizione di qualità proposta dalla ISO 8402, riferendola, però, al concetto di data item.

Una definizione di data item è:

“Dato un modello di dati, un data item è l’elemento a granularità più bassa del modello”

Esempi di data item sono gli attributi di una classe nel modello Object-Oriented (OO), o gli attributi di un’entità nel modello Entità-Relazione (ER).

Gli elementi del modello di dati costituiti da aggregazioni di data item, saranno genericamente riferiti come *entità*. Esempi di entità sono le classi del modello OO e le entità del modello ER.

La “qualità dei dati” può, dunque, essere definita come segue:

“Un insieme di caratteristiche che un data item deve avere affinché effettui i compiti per cui è stato concepito; tali caratteristiche sono specificatamente riferite ai valori ed al formato del data item”

La definizione introdotta è volutamente “limitativa”, ed in particolare non considera aspetti connessi allo schema concettuale dei dati,¹ concentrando l’attenzione sulla qualità dei valori e del formato.

Il motivo per cui si è introdotto il concetto di data item è connesso alla possibilità di un maggiore dettaglio nella definizione della *qualità dei dati*, rispetto alla definizione di *qualità* fornita dalla ISO 8402: mentre il concetto di entità nella definizione della ISO 8402 rimane volutamente generico, nel trattare la qualità dei dati si può far riferimento a qualcosa di più specifico, i data item appunto, inquadrandoli come gli elementi “base” (a granularità più fine) del modello dei dati sottostante.

Nei paragrafi successivi, sono individuate le dimensioni della qualità dei dati, mettendole in correlazione con le caratteristiche e sottocaratteristiche della qualità del prodotto software, descritte nello standard ISO/IEC 9126, ed adattate alle specificità del concetto di “dato”.

¹ Poiché non si considera la qualità di un data item come elemento del modello concettuale, nel corso del presente lavoro si useranno le espressioni “data item” e “valore del data item” in modo indifferente.

2.1 Le dimensioni della qualità dei dati

In letteratura, la qualità dei dati è un concetto ricondotto di solito ad un insieme di *dimensioni*. Spesso, le dimensioni considerate relativamente ad un data item sono l'accuratezza, la completezza o la attualità. Non esiste comunque un accordo su quali siano le dimensioni caratterizzanti la qualità dei dati, né sull'esatta definizione di quelle comunemente proposte. Ad esempio, la tempestività (in inglese "timeliness") è definita in termini di "disponibilità di dati in tempo" [9], oppure di "grado di aggiornamento dei dati" [2], quindi in termini di due concetti evidentemente differenti.

In [12], le dimensioni della qualità dei dati sono classificate in tre categorie: schema concettuale, valori e formato. Lo schema concettuale si riferisce all'aspetto "intensionale" dei dati, ovvero a quella componente dei dati che rimane invariata nel tempo. I valori sono invece relativi all'aspetto "estensionale" dei dati, ossia alla parte variabile nel tempo. Il formato è connesso alla modalità di rappresentazione dei dati.

Nel presente lavoro ci si focalizzerà sulle dimensioni della qualità dei dati connesse ai valori ed al formato, escludendo quelle legate allo schema concettuale. Il motivo di tale scelta è essenzialmente riconducibile ai tanti risultati raggiunti nell'ambito della ricerca sul concetto di qualità degli schemi concettuali ([3], [4], [13], [10], [6]), laddove, invece, la qualità dei valori e del formato si caratterizza per essere ancora ad uno stadio di evoluzione iniziale. Pur assumendo l'insieme delle dimensioni relative ai valori ed al formato proposto in [12], si è scelto di ridefinire e specializzare alcune delle dimensioni, di modo da risolvere parte delle ambiguità derivanti dal confronto con altre proposte. Ad esempio, come si dettaglierà in seguito, si è scelto di distinguere tra accuratezza temporale ed attualità per risolvere il conflitto connesso alle differenti definizioni di tempestività.

2.1.1 Le dimensioni della qualità dei valori

Le dimensioni individuate per la qualità dei valori sono:

- *Accuratezza sintattica*, definita come vicinanza del valore del data item ad un valore nel dominio di definizione *sintattico* considerato corretto.

Ad esempio, un basso grado di accuratezza sintattica corrisponde ad un valore di un data item che è Mrio anziché Mario.

- *Accuratezza semantica*, definita come vicinanza del valore del data item ad un valore nel dominio di definizione *semantico* considerato corretto.

Ad esempio, un basso grado di accuratezza semantica corrisponde al caso in cui si abbia Maria anziché Mario.

- *Accuratezza temporale* (in inglese "currency"), definita come vicinanza del valore del data item al valore temporalmente corretto. Si consideri, come esempio, il data item indirizzo di residenza; se Mario Rossi deve avere un indirizzo di residenza X alla data Dx e un indirizzo di residenza Y alla data Dy, qualora risulti che alla data Dy ha l'indirizzo X, il dato non è temporalmente corretto (anche se è corretto sintatticamente e "lo è stato" semanticamente).

- *Attualità* (in inglese "timeliness"), definita come adeguatezza del valore del data item rispetto alle esigenze temporali del contesto di utilizzo.

Se, ad esempio, l'orario delle lezioni relativo ai corsi di un'università è aggiornato "tardi" nella informativa destinata agli studenti, rispetto all'inizio dei corsi, il dato ha un livello insufficiente di attualità.

- *Completezza dei valori*, definita come l'estensione con cui i valori dei data item sono presenti nella base di dati. Un data item possiede un valore oppure è nullo; i valori nulli dovrebbero essere limitati, ed in ogni caso bisogna specificarne il significato². Questo tipo di completezza viene

² Si consideri un'entità Persona con attributo IndirizzoEmail; si possono avere i seguenti casi:

Valore non noto di un attributo applicabile (ad esempio si sa che la persona possiede un indirizzo email ma non è noto). E' un caso di incompletezza.

Attributo non applicabile (la persona non ha un indirizzo email). La completezza non è inficiata.

Il valore nullo è un valore valido per quell'attributo, fa cioè parte del dominio (nel contesto applicativo il valore nullo potrebbe indicare che l'indirizzo di email è riservato). È un indicatore di completezza.

talvolta chiamata completezza degli attributi. Si usa anche il termine completezza delle entità intendendo che tutte le istanze che dovrebbero essere presenti, effettivamente appaiono nella base di dati. Questa è più difficile da valutare³.

- *Consistenza Interna*, relativa al grado di consistenza di valori di data item nell'ambito della stessa entità

Ad esempio i valori dei data item CAP, Comune, Provincia, devono risultare consistenti nell'ambito dell'entità Persona.

- *Consistenza Esterna*, relativa al grado di consistenza di valori di data item nell'ambito di entità diverse.

Ad esempio, Mario Rossi è celibe in un'entità relativa al suo stato matrimoniale, mentre Anna Rossi si dichiara moglie dello stesso Mario Rossi, in un'entità di tipo diverso.

2.1.1.1 Relazioni tra alcune dimensioni della qualità dei valori

Le definizioni proposte nella precedente sezione suggeriscono forti analogie tra accuratezza sintattica e semantica e consistenza interna ed esterna. In particolare, in relazione a ciascuna di queste dimensioni, si può osservare che:

- l'accuratezza sintattica fa riferimento al dominio di definizione sintattico di un singolo data item nell'ambito della stessa entità
- l'accuratezza semantica fa riferimento al dominio di definizione semantico di un singolo data item nell'ambito della stessa entità
- la consistenza interna fa riferimento al dominio di definizione semantico di diversi data item nell'ambito della stessa entità
- la consistenza esterna fa riferimento al dominio di definizione semantico di diversi data item nell'ambito di entità diverse.

In Tabella 1 sono schematizzate le relazioni tra accuratezza semantica, consistenza interna e consistenza esterna.

	DOMINIO SEMANTICO	
	Singolo Data Item	Diversi Data Item
Singola Entità	AS	CI
Diverse Entità		CE

Tabella 1: Corrispondenza tra le dimensioni di accuratezza semantica, consistenza interna e consistenza esterna. Legenda: CI=Consistenza Interna; CE=Consistenza Esterna; AS=Accuratezza Semantica.

2.1.2 Le dimensioni della qualità del formato

Le dimensioni selezionate per la qualità del formato sono:

- *Usabilità*, è definita come appropriatezza del data item rispetto alle esigenze dell'utente; è influenzata dalle caratteristiche dell'utente e dal contesto tecnologico in cui il dato è utilizzato.

Ad esempio, i codici a barre sono maggiormente usabili dai dispositivi di lettura ottica che da utenti umani. Alcuni formati possono poi aiutare a risolvere problemi di accessibilità per i disabili.

- *Interpretabilità/Universalità del formato*, ha l'obiettivo di garantire una corretta interpretazione. Nell'accezione di universalità del formato, anche di garantire la comprensione a livello universale.

³ Si consideri il seguente esempio: nasce un bambino, i genitori non denunciano la cosa. Il database dell'anagrafe è incompleto (manca l'istanza che rappresenta il bambino) ma non c'è modo di risolvere la cosa (almeno a livello tecnologico, in realtà reingegnerizzando opportuni processi è possibile limitare questo problema).

Ad esempio, tra le rappresentazioni (1,2) e (maschio, femmina) è ovviamente più "interpretabile" la seconda.

Come esempio di universalità del formato, una rappresentazione con icone che corrispondono ad un uomo e una donna è maggiormente comprensibile, in un ambiente multilingue, rispetto ad una coppia di stringhe che in lingua italiana specificano il sesso (ad esempio maschio e femmina).

- *Precisione/Capacità di rappresentare valori nulli*, è relativa all'insieme di rappresentazione e rappresenta il grado con cui è possibile distinguere gli elementi di tale insieme rispetto alle esigenze dell'utente. Rappresentazioni precise devono anche garantire la possibilità di rappresentare valori nulli.

Ad esempio, si hanno diversi livelli di precisione a seconda che i valori numerici siano rappresentati con un formato integer oppure real.

- *Flessibilità*, è la facilità di cambiamento del data item in corrispondenza al cambiamento delle esigenze dell'utente od al cambiamento della piattaforma di supporto alla memorizzazione del data item.

Ad esempio, una rappresentazione di un data item anno a due cifre è stata sufficiente fino a quando non si è raggiunto l'anno 2000.

- *Efficienza di memorizzazione*, è relativa all'efficienza di utilizzo dei mezzi di memorizzazione.

Ad esempio, memorizzare due icone corrispondenti a maschio e femmina è meno efficiente che memorizzare (0,1); questo esempio mostra anche come alcuni requisiti possano essere in conflitto; infatti la memorizzazione di (0,1), pur essendo positiva per l'efficienza di memorizzazione, inficia la interpretabilità. E' dunque, in alcuni casi, necessario gestire un "trade-off" tra le varie dimensioni.

2.2 Lo standard ISO/IEC 9126

Nel 1991 l'ISO (International Organization for Standardization) e la IEC (International Electrotechnical Commission) hanno pubblicato la norma 9126 come standard internazionale. Lo standard definisce che la qualità di un prodotto software può essere rappresentata dall'insieme delle seguenti 6 caratteristiche: funzionalità, affidabilità, usabilità, efficienza, manutenibilità, portabilità.

Le caratteristiche, nella riedizione in via di pubblicazione [1], sono ulteriormente precisate attraverso 27 sottocaratteristiche, associate alle caratteristiche come mostrato in figura Figura 1.

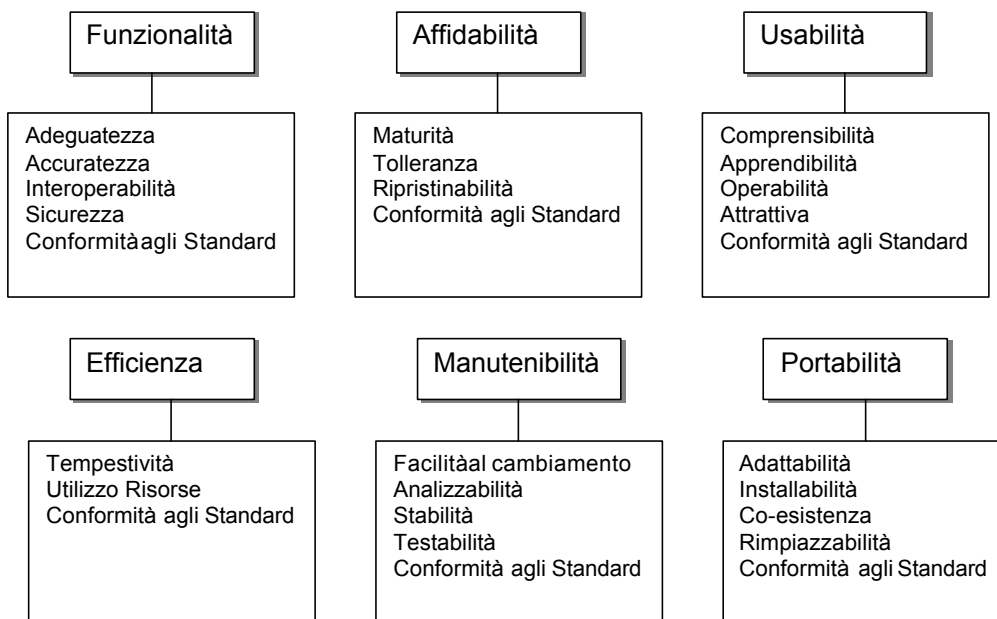


Figura 1: Le caratteristiche e le sottocaratteristiche dello standard ISO/IEC 9126.

E' inoltre effettuata la distinzione tra qualità interna, qualità esterna e qualità in uso: la qualità interna si riferisce alle caratteristiche intrinseche del prodotto software; la qualità esterna al comportamento del codice quando eseguito; la qualità in uso si riferisce alla capacità dell'utente di utilizzare con efficacia ed efficienza il prodotto.

Le caratteristiche e sottocaratteristiche, mostrate in Figura 1, rappresentano la qualità esterna ed interna, e sono quelle su cui è stata imperniata l'analisi critica ed il confronto con le dimensioni della qualità dei dati.

3. Corrispondenze tra le dimensioni della qualità dei dati e le sottocaratteristiche della ISO/IEC 9126

L'obiettivo di questo paragrafo è stabilire, laddove possibile, una corrispondenza tra le dimensioni della qualità dei dati e le sottocaratteristiche della ISO/IEC 9126. La corrispondenza è stabilita sulla base di un concetto intuitivo di similarità nelle definizioni delle dimensioni e delle sottocaratteristiche. Le sottosezioni 3.1 e 3.2 esaminano, rispettivamente, le dimensioni della qualità dei valori e del formato.

3.1 Le dimensioni della qualità dei valori e la ISO/IEC 9126

L'individuazione delle corrispondenze tra le sottocaratteristiche della 9126 e le dimensioni dei valori ha dato come risultato due casi di corrispondenza (vedi Tabella 2).

La prima corrispondenza riguarda una sottocaratteristica della funzionalità

- Accuratezza/Accuratezza Semantica

La seconda corrispondenza interessa una sottocaratteristica dell'efficienza:

- Tempestività/Attualità

Dimensioni Valori ----- caratteristica / sottocaratterist.	Accuratezza semantica	Attualità
Funzionalità/ Accuratezza	X	
Efficienza / Tempestività		X

Tabella 2: Corrispondenze ISO/IEC 9126 e dimensioni della qualità dei valori.

3.2 Le dimensioni della qualità del formato e la ISO/IEC 9126

Nel seguito si elencano le corrispondenze individuate, considerando le singole caratteristiche, e, per ciascuna sottocaratteristica, tracciando le corrispondenze con le dimensioni del formato.

Le sottocaratteristiche della funzionalità sono direttamente relazionabili ad alcune dimensioni della qualità dei dati associate al formato. In particolare definiamo le seguenti corrispondenze sottocaratteristiche della funzionalità-dimensioni del formato:

- Adeguatezza-Usabilità
- Accuratezza-Precisione/Capacità di rappresentare valori nulli

Le sottocaratteristiche dell'affidabilità non sono rapportabili in modo diretto alle dimensioni della qualità dei dati associate al formato.

Per l'usabilità è possibile individuare le seguenti corrispondenze:

- Comprensibilità/Interpretabilità/Universalità del formato
- Operabilità/Usabilità

Per la manutenibilità è stata individuata la seguente corrispondenza:

- Facilità al cambiamento-Flessibilità

Tra le sottocaratteristiche della efficienza, una corrispondenza diretta sottocaratteristica dell'efficienza-dimensione del formato è:

- Utilizzo risorse-Efficienza di memorizzazione

Tra le sottocaratteristiche della portabilità si individua la corrispondenza:

- Adattabilità-Flessibilità

Le corrispondenze individuate sono riassunte nella Tabella 3.

Dimensioni / formato ----- caratteristica / sottocaratterist.	Usabilità	Interpretabilità/ Univers. del formato	Precisione / Capacità di rapp. valori nulli	Flessibilità	Efficienza di mem.
Funzionalità/ Adeguatezza	X				
Funzionalità/ Accuratezza			X		
Usabilità/ Comprensibilità		X			
Usabilità/ Operabilità	X				
Manutenibilità/ Facilità di camb.				X	
Efficienza / utilizzo risorse					X
Portabilità/ Adattabilità				X	

Tabella 3: Corrispondenze ISO/IEC 9126 e dimensioni della qualità del formato.

3.3 Le relazioni tra le dimensioni e le sottocaratteristiche

La Tabella 4 riassume la corrispondenza, individuata nei precedenti paragrafi, tra le dimensioni della qualità dei dati e le sottocaratteristiche della 9126; questa corrispondenza induce le seguenti considerazioni:

- la corrispondenza non è di tipo *funzionale*: infatti, non tutte le dimensioni dei dati hanno almeno una sottocaratteristica della 9126 associata, ed alcune dimensioni possono essere associate a più di una caratteristica (si consideri, ad esempio la dimensione flessibilità del formato);
- tutte le dimensioni del formato hanno trovato una corrispondenza nelle sottocaratteristiche della 9126. In effetti ciò è giustificabile in base al fatto che il formato, ovvero la rappresentazione dei dati, è concettualmente legata al prodotto software: come il prodotto software “elabora” l'informazione per svolgere le funzioni del sistema informatico, così il formato “manipola” l'informazione per consentirne la memorizzazione a livello di sistema informatico;
- l'affidabilità è l'unica caratteristica della 9126 che non ha corrispondenze con le dimensioni della qualità dei dati. Infatti un dato non è affidabile (nel senso 9126 di tollerante ai guasti, ripristinabile, maturo ecc.) di per sé, ma l'affidabilità è, viceversa, una proprietà delle procedure/funzioni software che gestiscono il dato.

DIMENSIONI DELLA QUALITÀ DEI DATI		CARATTERISTICHE/SOTTOCARATTERISTICHE ISO/IEC 9126
Accuratezza Sintattica		Funzionalità/Adeguatezza
Accuratezza Semantica		Funzionalità/Accuratezza
Accuratezza Temporale		Funzionalità/Interoperabilità
Attualità		Funzionalità/Sicurezza
Completezza		Funzionalità/Conformità agli Standard
Consistenza Interna		Affidabilità/Maturità
Consistenza Esterna		Affidabilità/Tolleranza ai guasti
Usabilità		Affidabilità/Ripristinabilità
Interpretabilità/Univ. formato		Affidabilità/Conformità agli standard
Precisione/Cap. di rappr.e valori nulli		Usabilità/Comprensibilità
Flessibilità		Usabilità/Apprendibilità
Efficienza di memorizzazione		Usabilità/Operabilità
		Usabilità/Attrattiva
		Usabilità/Conformità agli standard
		Efficienza/ Tempestività
		Efficienza/Utilizzo Risorse
		Efficienza/Conformità agli standard
		Manutenibilità/Analizzabilità
		Manutenibilità/Facilità al cambiamento
		Manutenibilità/Stabilità
		Manutenibilità/Testabilità
		Manutenibilità/Conformità agli standard
		Portabilità/Adattabilità
		Portabilità/Installabilità
		Portabilità/Co-esistenza
		Portabilità/Rimpiazzabilità
		Portabilità/Conformità agli standard

Tabella 4: Tabella riassuntiva delle corrispondenze tra le dimensioni della qualità dei dati e le caratteristiche/sottocaratteristiche della ISO/IEC 9126.

4. Caso di studio

Nella presente sezione descriveremo le misure di qualità dei dati condotte sull'Archivio Anagrafico (AA) dell'Agenzia delle Entrate. La misurazione ha interessato le dimensioni dei valori e del formato descritte nella sezione 2.1; tali dimensioni sono il risultato di una selezione e sistematizzazione di quelle presentate in letteratura, finalizzata a poter effettuare una misurazione il più possibile esauriente e completa della qualità dei dati dell'AA.

4.1 Descrizione generale

La qualità dei dati che costituiscono il patrimonio informativo di un sistema riveste un ruolo cruciale per il sistema stesso; in particolare la necessità di avere a disposizione dati significativi ed attuali è un requisito che acquista un valore altamente strategico per l'Agenzia delle Entrate.

Nell'ambito dello sforzo di semplificazione degli adempimenti fiscali e dell'instaurazione di un nuovo rapporto tra l'Amministrazione Finanziaria e i contribuenti che preveda meno burocrazia, più servizi e maggiore trasparenza da parte dell'Amministrazione Pubblica, assumono sempre maggiore rilevanza le iniziative tese all'invio di comunicazioni ai contribuenti e di scambio dati con altre amministrazioni pubbliche. Il successo di tali iniziative dipende, in larga misura, dal livello di qualità del Sistema informativo dell'Agenzia.

Per questi motivi l'Agenzia delle Entrate si è proposta di attuare un Piano di Qualità dei Dati che consta di tre fasi principali e parallele:

1. *Misurazione della qualità dei propri dati*, data cleaning e realizzazione di strumenti per il monitoraggio e il mantenimento continuo della qualità
2. *Riorganizzazione delle banche dati*. Tale fase riguarda in particolare le basi di dati reddituali e anagrafiche. Oltre ad un nuovo disegno delle basi dati, la riorganizzazione prevede: (i) l'individuazione delle fonti di replicazione e di ambiguità delle informazioni presenti nella vecchia struttura delle basi dati; (ii) la risoluzione di tali ambiguità nel nuovo disegno, attraverso l'introduzione di nuovi campi. Un esempio di ambiguità nella definizione dei campi è fornito da un campo "data di variazione", che può assumere diversi significati a seconda che sia riferito alla data in cui un certo evento si è verificato ovvero a quella in cui tale evento è stato registrato nel sistema informatico; una tale ambiguità può essere risolta introducendo nuovi campi, del tipo "data di aggiornamento" dell'evento e "data di comunicazione" dell'evento. Il processo di riorganizzazione è iniziato con le basi dati reddituali a seguito dell'introduzione, nel 1998, della dichiarazione unica e dell'introduzione della presentazione per via telematica della stessa, ed è tuttora in corso con riferimento alle basi dati anagrafiche.
3. *Revisione dei processi*. Il mantenimento della qualità dei dati a lungo termine può essere conseguito, mediante un'analisi dei processi e un'individuazione delle loro criticità nella produzione di dati di qualità. A tal fine si è iniziato a distribuire ai cittadini, agli intermediari fiscali e agli enti esterni all'Amministrazione Finanziaria (quali Comuni, Ordini professionali e altre Pubbliche Amministrazioni) prodotti per la predisposizione e l'invio telematico delle dichiarazioni, degli atti e delle comunicazioni all'Agenzia delle Entrate, in modo da consentire il controllo della qualità dei dati tempestivamente e di localizzare dove questi ultimi vengono generati. Inoltre si stanno introducendo strumenti per il monitoraggio dei flussi di dati (costituzione di un'anagrafe dei fornitori, tipologia di informazioni fornite da ciascun ente, quantità di informazioni fornite, modalità di trasmissione utilizzata, ecc.).

Le tre fasi sono connesse tra loro e si influenzano e alimentano reciprocamente secondo lo schema riportato in Figura 2.

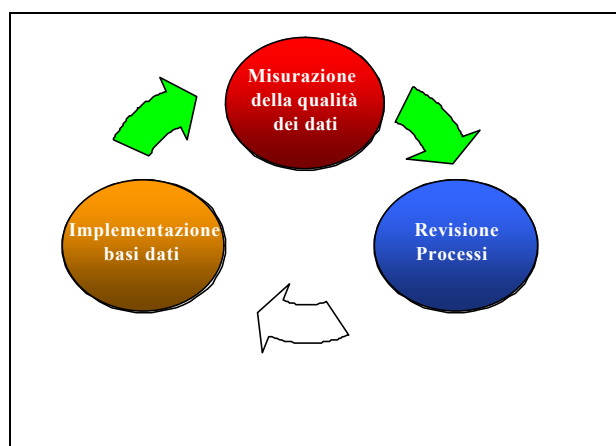


Figura 2: Le fasi del Piano di Qualità dei dati dell'Agenzia delle Entrate.

Il caso di studio descritto nella presente sezione interessa la prima delle fasi elencate, ossia la fase di misurazione della qualità dei dati. In particolare l'attività di misurazione è stata condotta sull'Archivio Anagrafico (AA) delle persone fisiche, in virtù del ruolo cruciale che questo riveste nell'attività dell'Agenzia delle Entrate, per la quale è fondamentale l'individuazione certa del cittadino e la possibilità di rintracciarlo correttamente e tempestivamente.

4.2 Le dimensioni misurate

I campi degli archivi della AA sui quali sono state condotte misure di qualità sono indicati in Figura 3.

Codice Fiscale	Cognome	Nome	ComuneNascita	DataNascita	ComuneResidenza	Indirizzo	Informazioni Decesso
----------------	---------	------	---------------	-------------	-----------------	-----------	----------------------

Figura 3: Campi dell'AA.

Allo stato, le misure di qualità per ciascun campo sono state condotte solo per un sottoinsieme delle dimensioni di qualità dei dati elencate nelle precedenti sezioni: le dimensioni dei valori sono state misurate quasi completamente (l'unica eccezione è l'accuratezza temporale), mentre, per quanto riguarda quelle del formato la misurazione è in una fase ancora iniziale.

In particolare, le dimensioni relative ai valori che sono state misurate sono:

- Accuratezza sintattica
- Accuratezza semantica
- Attualità
- Completezza
- Consistenza interna
- Consistenza esterna

Per quanto riguarda le dimensioni del formato, è stata misurata la sola:

- Flessibilità

Tra le dimensioni misurate, quelle messe in corrispondenza con le sottocaratteristiche della ISO 9126 sono l'accuratezza semantica, l'attualità e la flessibilità. L'accuratezza semantica è stata associata alla sottocaratteristica della funzionalità costituita dall'accuratezza; l'attualità è stata associata alla tempestività, sottocaratteristica dell'efficienza; la flessibilità è stata associata alla flessibilità al cambiamento, sottocaratteristica della manutenibilità. Per la misura della attualità si sono mutate tecniche derivanti dalla corrispondenza con la ISO 9126 (su cui si entrerà nel dettaglio successivamente).

Nel seguito saranno riportate le tecniche di misura adottate per ciascuna dimensione ed i risultati di misura ottenuti.

Accuratezza Sintattica e Semantica

L'accuratezza sintattica e l'accuratezza semantica sono state misurate mediante *confronto con dizionari sintattici e semantici di riferimento*.

I campi interessati da tale misurazione sono i campi Nome e Cognome, ed è stato analizzato un campione di 596.903 anagrafiche. È stato adottato un procedimento congiunto di riconoscimento e correzione, e sono stati distinti quattro possibili casi:

- Anagrafica riconosciuta senza modifiche, per la quale sono stati individuati e risolti i “match” con i dizionari di riferimento.
- Anagrafica riconosciuta con modifiche, per la quale i match sono stati risolti a valle di modifiche. Ad esempio sono stati individuati e corretti errori sintattici del tipo Giovanni in Giovanni.
- Anagrafica ambigua, per la quale sono stati individuati ma non risolti i possibili match.
- Anagrafica non riconosciuta, per la quale non sono stati individuati dei match con i dizionari di riferimento.

I risultati della misurazione sono riassunti in Tabella 5.

	Numero Record	%
Anagrafica riconosciuta senza modifiche	589.356	98,74%
Anagrafica riconosciuta con modifiche	3.412	0,57%
Anagrafica ambigua	1.316	0,22%
Anagrafica non riconosciuta	2.819	0,47%
Totale	596.903	100,00%

Tabella 5: Risultati relativi alle misure di accuratezza sull'anagrafica.

Completezza

La completezza è stata misurata mediante *verifica ispettiva della presenza dei valori* dello specifico campo sull'intero Archivio Anagrafico e confronto con requisiti di specifica.

I campi per i quali è stata misurata la completezza sono: ComuneNascita, ComuneResidenza, Indirizzo; InformazioniDecesso. I risultati sono riassunti nella Tabella 6.

DimensioneQualità ----- Tipo Dato	Requisito Completezza	Indicatore di completezza
ComuneNascita	Sempre presente	100%
ComuneResidenza	Sempre presente	100%
Indirizzo	Sempre presente	99,8%
Informazioni Decesso	Data e fonte sempre presente	92,8%

Tabella 6: Risultati relativi alle misure di completezza.

Consistenza Interna

La consistenza interna è stata misurata mediante *applicazione di business rules*⁴, ed è stata misurata sull'intero AA.

Ad esempio il campo Codice Fiscale è stato confrontato con i campi Nome, Cognome, Data di Nascita, in base alla dipendenza dell'algoritmo di calcolo del codice fiscale da tali valori; tale dipendenza costituisce una “business rule”.

Un'ulteriore business rule che è stata verificata, è relativa al campo InformazioniDecesso rispetto a quello Data di Nascita, in cui ovviamente deve risultare posteriore la data del decesso rispetto a quella di nascita. I risultati sono riportati nella Tabella 7.

⁴ Le business rules (vincoli o data edits) sono routine computerizzate che verificano se i valori soddisfano vincoli predeterminati. I valori non corretti possono essere corretti o in base ai vincoli stessi o facendo riferimento al mondo reale.

DimensioneQualità ----- Tipo Dato	Requisito Consistenza Interna	Indicatore di Consistenza Interna
Codice Fiscale	Coerente con i dati anagrafici	100%
Informazioni decesso	Coerente con data di nascita	99,55%

Tabella 7: Risultati relativi alle misure di consistenza interna.

Consistenza Esterna

La consistenza esterna è stata misurata mediante *confronto con basi di dati di riferimento*⁵. In particolare le basi di dati utilizzate sono state quelle dei comuni di residenza. La misurazione è stata effettuata su tutti i campi (vedi Tabella 8).

Tipo Dato ----- Dimensione Qualità	Cod.Fiscale	Cognome Nome	Comune Nascita	Data Nascita	Comune Residenza	Indirizzo	Informazioni Decesso
Consistenza Esterna	Confermato dal Comune di residenza Attribuito dal Comune	Confermati dal Comune di residenza	Confermati dal Comune di residenza	Confermati dal Comune di residenza	Confermati dal Comune di residenza	Confermati dal Comune di residenza	Confermati dal Comune
Indicatore	87%(1) 25,5% (2) 17% (3)	87%(1) 25,5% (2)	87% (1) 25,5% (2)	87% (1) 25,5% (2)	87% (1) 25,5% (2)	87% (1) 25,5% (2)	71% (1) 13,5% (2)

Tabella 8: Risultati relativi alle misure di consistenza esterna. (1)= Percentuale calcolata su un campione di 22.167.345 soggetti confrontato con i dati dei Comuni; (2)= Percentuale calcolata sull'intera popolazione dell'AA.; (3)= Percentuale calcolata sui codici fiscali attribuiti dal comune di residenza per i nati nell'anno 2000 (612.941 soggetti).

Flessibilità

La flessibilità è stata misurata per un campo particolare, il campo Indirizzo, sulla base della "normalizzabilità" di tale campo, ovvero sulla base della *possibilità di ricondurre il campo ad un formato normalizzato di riferimento*. I risultati di misurazione su un campione di 275.365 indirizzi sono riassunti nella Tabella 9.

DimensioneQualità ----- Tipo Dato	Requisito di Flessibilità	Indicatore di Flessibilità
Indirizzo	Normalizzabile	80,9%

Tabella 9: Risultati relativi alle misure di flessibilità

Attualità

Una delle possibili applicazioni della corrispondenza, riassunta nella Tabella 4, tra le dimensioni dei dati e le sottocaratteristiche di qualità con siste nell'utilizzare i criteri (ovvero le proprietà misurabili delle caratteristiche di qualità), definiti in generale per la misura della qualità di un prodotto software, adattandoli al caso della misura della qualità dei dati.

Nella Tabella 10 si riporta un esempio di metrica associata alla dimensione dei valori costituita dalla attualità ed adattata a partire dai criteri della ISO 9126.

⁵ Questa tecnica è nota in letteratura come database bashing (comparare i dati tra due o più database). In questo modo vengono comparati i record di database differenti: se coincidono si assumono corretti, se discordano vengono entrambi controllati ed eventualmente corretti (ad esempio facendo riferimento al mondo reale).

Dimensione	Metrica	Finalità della Metrica	Metodo di applicazione	Elementi di misurazione	Formula ed interpretazione del valore misurato	Tipo della Scala di misurazione	Input alla misurazione
Attualità	Attualità dello specifico data item	Quanto è attuale il data item rispetto ai requisiti specificati?	Rilevamento dei tempi di risposta dei processi interessati alla produzione/gestione del data item	$TR_i = \text{Tempo di risposta dell'i-esimo processo rilevato}$ $TR = \sum TR_i$ $R_{max} = \text{Max ritardo accettabile atteso}$	$TR \leq R_{max}$	Assoluta	Specificazione del requisito R_{max}

Tabella 10: Una metrica per la misura dell'attualità ricavata a partire dai criteri della 9126.

I passi metodologici seguiti nella misurazione dell'attualità sono descritti in seguito, e verranno illustrati mediante l'esempio del processo seguito per il calcolo dell'attualità relativa al data item Codic e Fiscale. Si osserva che le misure condotte hanno carattere esemplificativo: quando sarà condotta la fase di revisione dei processi prevista dal Piano di Qualità dei Dati dell'Agenzia delle Entrate, le misure di attualità potranno essere derivate in maniera più puntuale e completa, potendo riferirsi ad un'analisi di dettaglio dei processi e del flusso dei dati.

I° passo: Rilevamento dei processi interessati alla produzione/gestione dei data item.

Il data item che si considera è il Codice Fiscale.

Il processo di attribuzione del codice fiscale è descritto in Figura 4. Il Comune comunica all'Agenzia delle Entrate i dati anagrafici di un individuo; a questo punto viene effettuata l'attribuzione del codice fiscale in maniera integrata o manuale a seconda della disponibilità di procedure integrate da parte del Comune.

II° Passo: Specifica dei valori R_{max} per ciascuno dei processi individuati.

Nell'effettuare le misurazioni si è considerato un unico processo di attribuzione dei codici fiscali, essendo in una fase iniziale della misurazione. Come accennato, nella fase di revisione dei processi il dettaglio dell'analisi sarà maggiore, al fine di poter individuare tutte le possibili criticità un esempio è però rilevabile anche in Figura 4, in cui il ramo di processo relativo al data entry manuale, costituisce certamente una criticità

Il valore di ritardo massimo (R_{max}) stabilito per il processo di attribuzione dei codici fiscali è:

$$R_{max} = \text{quattro mesi dalla data di nascita}$$

III° passo: Calcolo dei TR_i e di TR per ciascuno dei processi individuati.

Nel nostro esempio $TR = TR_1$ essendo stato considerato un unico processo.

Il valore TR calcolato è:

$TR = 293$ gg nel 2000

$TR = 82$ gg nel 2001 (I semestre)

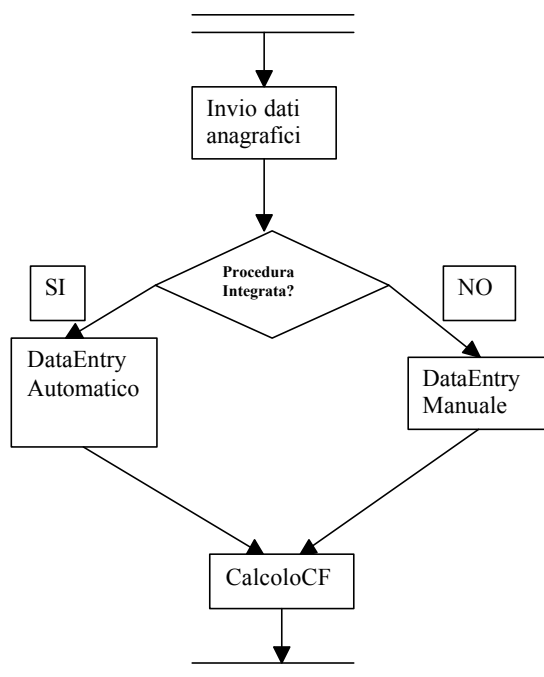


Figura 4: Il processo di acquisizione dei codici fiscali.

IV° passo: Interpretazione della formula di misura

Nell'esempio si ha:

$TR=293gg > R_{max}=120gg$ nel 2000
 $TR=82gg < R_{max}=120gg$ nel 2001 (I semestre)

Ne consegue una violazione dei requisiti di specifica del tempo massimo, anche consistente (173 gg pari al 144% di R_{max}) per l'anno 2000.

Di contro, il rispetto dei requisiti di specifica del tempo massimo per il primo semestre dell'anno 2001 è abbastanza contenuto (38 gg pari al 32% di R_{max}).

5. Conclusioni e Sviluppi Futuri

Nel presente lavoro è stato individuato un insieme di dimensioni di qualità dei dati, relative ai valori ed al formato dei dati, per il quale: (i) è stata stabilita una corrispondenza con le sottocaratteristiche della ISO 9126; (ii) è stata effettuata una misurazione sui dati del sistema informativo dell'Agenzia delle Entrate. La corrispondenza "dimensioni qualità dei dati -sottocaratteristiche ISO 9126" è stata inoltre sfruttata per definire una metrica per l'attualità a partire da quella definita per la corrispondente sottocaratteristica della tempestività.

Uno sviluppo futuro consiste nell'estensione delle misure effettuate a campione all'intera popolazione dell'AA. Inoltre si vogliono indagare ulteriori metriche per le dimensioni della qualità dei dati, definibili a partire dalla corrispondenza con lo standard ISO 9126. Un altro obiettivo futuro è quello di formalizzare le definizioni di qualità dei dati, al fine di poterle assumere come un insieme di riferimento esattamente "interpretabile".

6. Bibliografia

- [1] Azuma M.: Square, the Next Generation of the ISO/IEC 9126 and 14598 International Standard Series on Software Product Quality. In Proceedings of the 12th European Software Control and Metrics Conference, London, 2001.
- [2] Ballou, D. P. and Pazer H. L., Modeling Data and Process Quality in Multi-input, Multi-output Information Systems. Management Science, vol. 31, no 2, 1985.
- [3] Batini C., Lenzerini M., Navathe S.B. (1984): A comparative analysis of methodologies for database schema integration. ACM Computing Survey, vol. 15, no. 4, 1984.
- [4] Calvanese D., De Giacomo G., Lenzerini M., Nardi D., Rosati R.: Information Integration: Conceptual Modeling and Reasoning Support. In Proceedings of the 6th International Conference on Cooperative Information Systems (CooplS'98), New York City, NY, USA, 1998.
- [5] Crosby P.B.: Quality without Tears. McGraw-Hill, New York, 1984.
- [6] Gertz M.: Managing Data Quality and Integrity in Federated Databases. Second Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems, Airlie Center, Warrenton, Virginia, 1998.
- [7] ISO 8402: Quality Management and Quality Assurance-Vocabulary.
- [8] ISO/IEC 9126: Information Technology- Software Product Evaluation- Quality Characteristics and Guidelines for their Use.
- [9] Kriebel, C. H. and O. Mikhail, Dynamic Pricing of Resources in Computer Networks. Logistics, 1980.
- [10] Madnick S.: Metadata Jones and the Tower of Babel: The Challenge of Large -Scale Semantic Heterogeneity. Proceeding of the 3rd IEEE Meta-Data Conference (Meta-Data '99), Bethesda, MA, USA, 1999.
- [11] Natale D.: Qualità e Quantità nei Sistemi Software. Franco Angeli, 1995.
- [12] Redman T.C.: Data Quality for the Information Age. Artech House, 1996.
- [13] Ulmann J.D.: Information Integration using Logical Views. Proceedings of the International Conference on Database Theory (ICDT '97), Greece, 1997.
- [14] Wang R.Y., Strong D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, vol. 12, no. 4, 1996.